THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Computational methods for analysis of fragmented sequence data

Fredrik Boulund

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2015

Computational methods for analysis of fragmented sequence data
Fredrik Boulund
Göteborg 2015
ISBN 978-91-7597-281-7

Doktorshavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3962
ISSN 0346-718X

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX
Printed by Chalmers Reproservice
Göteborg, Sweden 2015

Computational methods for analysis of fragmented sequence data

Fredrik Boulund

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Recent developments in genomic and proteomic sequencing technologies have revolutionized research in life sciences, providing new opportunities for the study of biological systems. However, modern sequence data sets are large, diverse, and heavily fragmented, which presents new challenges for their analysis and interpretation. In this thesis we present six research papers, that describe novel methods for studying bacteria and bacterial communities through the analysis of large data sets produced by modern DNA and protein sequencing technologies.

In Paper I, we describe a method for discovering fragments of fluoroquinolone antibiotic resistance genes in short fragments of DNA. The resistance phenotypes of the predicted resistance genes were then validated by expression in an *Escherichia coli* host (Paper II). The method was further improved to handle larger and more fragmented data sets in Paper III. In Paper IV, we present Tentacle, an easy-to-use tool for high performance gene quantification in metagenomes that can be run on distributed computing resources to enable fast and efficient gene quantification in terabase metagenomes. In Paper V, we introduce proteotyping, an approach for microbial identification in clinical samples based on shotgun proteomics. Finally, in Paper VI we describe and evaluate a method for proteotyping analysis suited for application to clinical diagnostics of bacterial infections.

The rapidly increasing volumes of data produced by new sequencing technologies provide new opportunities for understanding microbial biology. To unlock the full potential of large sequence data sets requires novel methods and approaches such as those presented in this thesis.

Keywords: bioinformatics, metagenomics, proteomics, sequencing, distributed computing.

List of publications

This thesis is based on the work represented by the following papers:

I. **Boulund, F.**, Johnning, A., Pereira, M.B., Larsson, D.G.J., Kristiansson, E. (2012). A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. *BMC Genomics, **13**:695*, doi: 10.1186/1471-2164-13-695.

II. Flach, C-F., **Boulund, F.**, Kristiansson, E., Larsson, D.G.J. (2013). Functional verification of computationally predicted qnr genes. *Annals of clinical microbiology and antimicrobials*, **12.1**:34, doi: 10.1186/1476-0711-12-34.

III. **Boulund, F.**, Berglund, F., Bengtsson-Palme, J., Flach, C-F., Larsson, D.G.J., Kristiansson, E. (2015). Computational prediction of novel fluoroquinolone antibiotic resistance genes in public metagenomic data sets. *Manuscript*.

IV. **Boulund, F.**, Sjögren, A., Kristiansson, E. (2015). Tentacle: distributed quantification of genes in metagenomes. *GigaScience*, **4**:40, doi: 10.1186/s13742-015-0078-1.

V. Karlsson, R. Siles, L.G., **Boulund, F.**, Lindgren, Å. Svensson-Stadler, Å., Karlsson, A., Kristiansson, E., Moore, E.R.B. (2015). Proteotyping: Tandem Mass Spectrometry and Proteomic Analysis of Pathogenic Microorganisms. *Manuscript*.

VI. **Boulund, F.**, Karlsson, R., Siles, L.G., AL-Bayati, O., Moore, E.R.B., Kristiansson, E. (2015). A computational method analyzing mass spectrometry proteomics data for bacterial identification and abundance estimation suitable for pure cultures and mixed samples. *Manuscript*.

Additional papers not included in this thesis:

VII. Karlsson, R., Siles, L.G., **Boulund, F.**, Svensson-Stadler, L., Skovbjerg, S., Karlsson, A., Davidsson, M., Hulth, S., Kristiansson, E., Moore, E.R.B. (2015). Proteotyping: Proteomic characterisation, classification and identification of microorganisms — a prospectus. *Systematic and Applied Microbiology*, **38**:4, doi: 10.1016/j.syapm.2015.03.006.

VIII. Bengtsson-Palme, J., **Boulund, F.**, Weijdegård, B., Flach, C-F., Fick, J., Kristiansson, E., Larsson, D.G.J. (2014). Metagenomics reveal a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Frontiers in Microbiology*, **5**:648, doi: 10.3389/fmicb.2014.00648.

IX. Hakimi, C.S., Hesse, C., Wallén, H., **Boulund, F.**, Grahn, A., Jepsson, A. (2015). In vitro assessment of platelet concentrates with multiple electrode aggregometry. *Platelets*, **26**:2, pp 132-137. doi: 10.3109/09537104.2014.898141

Author contributions

I. Participated in study design, developed and wrote the implementation of the method, manually curated the alignment used to construct the HMM, optimized classification parameters for HMM matches, performed data analysis and semi-manual curation of fragment assemblies, wrote the online documentation, drafted and edited the manuscript.

II. Participated in study design, assembled the nucleotide and amino acid sequences for the candidate genes, and edited the manuscript.

III. Participated in study design, selected the data sets, participated in pre- and post-processing of the data, analyzed the identified sequences and clusters, constructed the gene tree, drafted and edited the manuscript.

IV. Participated in study design. Developed, implemented, and optimized the bioinformatics workflow. Contributed to the design and development of the distribution framework. Designed and executed the performance evaluation of the method on high-performance computing systems, wrote the online documentation, drafted and edited the manuscript.

V. Performed the bioinformatics analyses, wrote the sections on bioinformatics and the relevant algorithms, and described the example implementation.

VI. Participated in study design. Developed and implemented the overall workflow, optimized parameters settings. Performed the bioinformatic analysis of the *in silico* and experimental evaluations of the method. Wrote the documentation and drafted and edited the manuscript.

Acknowledgements

There are many people that have helped me on my way (and sometimes helped me find what way that was in the first place). **Erik Kristiansson**, supervisor and nerdy friend. Thank you so much for always being there and having the almost unreal quality of always making me see things clearly and making me feel better in any situation. I can't imagine a better supervisor. Thank you co-supervisors **Anders Sjögren** and **Marija Cvijovic** for your excellent support and continuous helpfulness. I've learned a lot from you; thanks for all discussions on both personal and professional development. **Olle Nerman**, without you I wouldn't have gone the bioinformatics route to begin with. I'm sometimes awestruck by your ability to always provide great and on-point feedback.

I've had the absolute pleasure to get to know a whole lot of amazing people over the past few years. **Anna Johnning**, you've been there all along, supporting and helping me in both science, and life in general, whenever I needed it. You've given me memories playing games, sharing music, traveling, and just having fun that I'm never going to forget. **Viktor Jonsson**, my buddy since day zero. To imagine that we would make it this far together when we first met more than nine years ago. I hope we can continue our nerdy relationship for years to come (oh, how fondly I remember those SupCom-games). **Johan Palme**, you and I have been walking side-by-side for a while now and it's been astounding! You have a solid position in the geek-squad and your skills in science and music are inspiring. I really enjoyed all the time we spent together both in and out of work. **Kaisa Thorell**, your happy outlook on life and aspirations for mastering everything you come across are inspiring and I'm overjoyed to have met you. Collaborating with you, be it climbing, conference organization, or hiking, is always great. Let's continue doing that. **Oskar Hamlet**, I really hope we can share more experiences like our Jotunheimen-trip together in the future. I promise I'll brush up on my crossword skills until next time. **Erik Sterner**, maybe you're actually the reason I thought a science career could be fun–our time in *the super group* has formed the foundation of what I think a good research team should be like. **Leif Väremo**, I won't forget the times we spent together bouldering on Hönö or developing image analysis algorithms. **Franscesco Gatto**, I can't imagine anyone more positive than you. It's been fun working with you on the organization of SBW2014, climbing, and just hanging out with you in general. All the people connected to "Kristiansson group" at Chalmers contribute to why it is so much fun to go to work every day. You are all great to hang around to discuss both serious and geeky stuff with: **Anna Rehammar, Mariana Pereira, Fanny Berglund, Tobias Österlund, Emma Wijkmark**, and **Marina Axelson-Fisk**.

I also learned a lot from the people I got to know during my time at Sahlgrenska Academy. Thank you. **Joakim Larsson**, for being an inspiring group leader. **Filip Cuklev Stern**, for discussions on music and wonderful climbing excursions. **Lina Gunnarsson-Kearney**, for your unfaltering positivism and for introducing me to bouldering. **Carolin Rutgersson**, always looking at the bright side of life. **Carl-Fredrik Flach**, for insightful discussions on science and the fun we've had (vad ska du ha på dig ikväll?). **Ida Nilsson**, the perpetually smiling room mate. Juntan is what made me realize that pursuing a research career in academia could actually be fun! I'm also really happy for all our Friday bouldering sessions **Robert Jakubowicz**, always taunting me to keep improving my climbing ("spänn magen"). The time with you guys taught me about research life in general, that poop is an acceptable lunch topic, that climbing is fun, and your special humor just inspire me to never accept not having fun in the workplace.

There is a long list of great people at the department that I always feel I never get to see enough of: **Henrike, Magnus Önheim, Peter, Dawan, Claes, Alexey, Ivar, Roza, Malin, Matteo, Magnus Röding, Tobias A, Johannes, Jonas, Jonathan, José, and everyone else that I missed**. Thank you for being the reason why it was so easy to spend a little too much time in the lunchroom :).

I really want to thank all my collaborators that I've met and worked with over the years. It is my firm belief that collaboration creates something greater than the sum of its parts, and working with you has reinforced that belief. Special mentions to **Roger Karlsson**, **Lucia Gonzales-Siles**, and **Edward Moore**.

Thanks to **Marianne Rosander-Bäckström**, **Cecilia Gelin**, **Lotta Fernström**, **Marie Kühn**, and **Jovan Pankovski**, for helping me out with all the administrative stuff I didn't know anything about. You were always so helpful and friendly, despite me asking the same questions over and over again.

I owe a great deal of my sanity to my best friends and fellow band members in DÖDAREN: **Jon Solheim, Martin Claesson, Eric Rådegård**, with whom I've had a blast making both happy memories and awesome music. I can't imagine what the past five years would've been without you. Spending time with you was the best creative outlet I could've had when I needed to focus on something else. Thanks to **Josefin Westin** and **Ida Rådegård** for being part of the extended DÖDAREN-family, I've had some great times together with you.

I'm very grateful for my family and your ability of always listening to me even if the only thing that's perfectly clear is that I'm not really making much sense. Thank you **Dad** for making me interested in science by doing experiments in the kitchen with me as a kid. Thank you **Mom** for pushing me to always strive for my goals. And thank all of you my great sisters **Clara, Vera**, and **Ulrika**.

My lovely wife, **Sara**. Thank you for being you and for your unwavering support. Sharing my life with you is fantastic and I would marry you again in heartbeat.

Contents

1 Introduction

Bioinformatics is a field of research at the intersection of biology, computer science, and mathematics. Biological research strives to understand the functions and processes in microorganisms, ourselves, and diseases to understand how these functions and processes work and interact. In bioinformatics research we develop tools and methods for untangling and making sense of biological data. Often, raw biological data are large and complex, making it difficult to directly understand the biological processes being studied. The computer science component of bioinformatics contributes with knowledge and methodology for the development of suitable and efficient algorithms, software, and techniques to analyze the sometimes massively large digital representations of biological data sets. Mathematics and statistics are essential components used for making inferences and predictions based on biological data, producing models describing the underlying biological processes and their variability so that hypotheses can be formulated. These hypotheses can, in turn, be investigated with experiments, thereby increasing our understanding of these biological systems. In addition, bioinformatics is also used as a term in reference to the computational tools and algorithms developed to extract useful information from biological data.

This thesis is based on six papers, which cover several topics mostly centered on the analysis of fragmented genomic and proteomic data. Applications include the identification of novel antibiotic resistance genes in the environment, utilizing high performance computing clusters to quantify gene content in microbial communities, and extracting information regarding the microbial contents of a sample based on the expressed proteins present in the sample. Because of the amalgamation of disciplines, methods, and analysis approaches, much conceptual ground must be covered to be able to fully appreciate the challenges inherent to bioinformatics research. Thus, before presenting the papers, the thesis begins with an overall introduction that is divided into five chapters. First, the background chapter introduces the necessary biological

concepts, sequencing technologies, and challenges arising from the application of these technologies to the problem of answering biological questions. The second chapter presents the main aims that instigated the work described in the six papers, and connects the aims to the findings in each paper. Chapter three presents a complement to the methods sections of the individual papers. Chapter three expands on some of the background to the methodological approaches taken, and provides a brief overview of how the analyzed data is stored in the computer. The fourth chapter summarizes the six papers, highlighting key concepts, results, and contributions to the field brought forth by the research underlying the papers. Chapter five presents conclusions of the research in the papers and discuss future perspectives.

2 Background

2.1 Metagenomics

Bacteria are single cell microbial organisms that come in different shapes and sizes, and are primarily named by their shape, e.g., bacillus (rod-shaped), coccus (spiral), and vibrio (comma-shaped). These unicellular organisms reproduce asexually and generally exhibit high reproduction rates. Some species of bacteria are capable of doubling in number every 20 minutes. Reproduction occurs by splitting each cell into two, meaning that a single bacterium can rapidly transform into a large number of bacteria. Bacteria often thrive and cooperate in complex communities. It has been demonstrated that many bacteria can act unselfishly in some situations, working together for the greater good of the whole community, rather than acting as single disconnected cells trying to outcompete all of their neighbors (Jefferson, 2004; Kreft, 2004). Bacteria are practically omnipresent and have found ecological niches that sustain them in almost every habitat imaginable, and have been observed dwelling in the deepest parts of the ocean, in hot springs, frozen in arctic ice, in soil, inside other organisms, and in the air high up in the atmosphere. In their seminal paper, Whitman et al. (1998) produced the first real estimate of the total number of bacteria on Earth. According to their estimate, the total number of bacteria on Earth could be around 5×10^{30} cells (Whitman et al., 1998). That figure is five *nonillion* cells on the short scale used in most English speaking regions, equivalent to five *quintillion* on the long scale normally used in Sweden.

Most people probably have mainly negative associations with bacteria, connecting them with various diseases and sickness. However, the majority of the bacterial organisms that we share our planet with are not pathogenic to us. The total number of bacterial species has been estimated to more than 10 million (Schloss and Handelsman, 2004), and only around five hundred of these (0.005%) are pathogenic to humans (Taylor et al., 2001). In fact, hu-

mans are largely dependent on, and colonized, by a wide range of bacteria. A recent paper estimated the total number of human cells in a human body to approximately 4×10^{13} cells (Bianconi et al., 2013), and the average human body contains about ten times more bacterial cells than human cells. The composition of the human microbiome and its interplay with our own systems are believed to have a profound effect on our health (Hooper and Gordon, 2001). The realization that bacteria play a big role in human health has spurred a wealth of large efforts to elucidate the interactions between our bacterial communities and ourselves (Qin et al., 2010; Huttenhower et al., 2012).

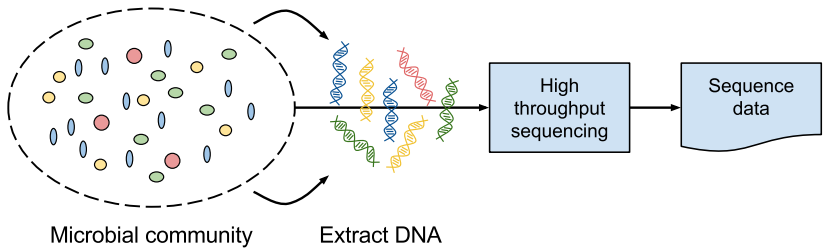


Figure 2.1: An overview of the shotgun metagenomics approach to studying microbial communities. Depicted to the left is a microbial community with organisms (e.g. bacteria and viruses) represented by small colored ellipses and circles. DNA from organisms in the community is randomly sampled and enzymatically fragmented before being sequenced using a high throughput sequencing machine. The resulting sequence data corresponds to random fragments of DNA from all the genomes of all the organisms in the community.

The study of bacterial communities is not straightforward. Historically, learning about a bacterial species relied on the ability to produce a pure culture of this species, so that its properties could be studied in isolation. However, this approach is quite problematic when studying complex communities where most species of bacteria are difficult to culture using traditional methods. For example, studies estimate that up to 99% of all environmental bacteria are difficult to grow in the lab using culture-based methods (Hugenholtz et al., 1998; Stewart, 2012). Instead of growing bacteria in the lab to study them, metagenomics can be used. Metagenomics is a technique for examining microbial communities by studying their genetic content directly. Analogous to how the term genome describes the complete genetic makeup of a single organism, a metagenome is the collective genetic material of an entire community of organisms (which might include eukaryotes, archaea, and viruses). The term was originally introduced by Handelsman et al. (1998) to describe the application of genomic techniques to complex mixtures of microorganisms. Figure 2.1 shows an overview of how metagenomics is used to study microbial communities. DNA is first extracted from a microbial community of interest. The DNA is

fragmented using enzymes that cut the DNA strands into shorter pieces. The DNA fragments are then put into a high throughput sequencing machine that determines the ordered sequence of nucleotides in the sampled fragments.

The concept behind metagenomics is well suited for investigating microbial communities. One of the first uses of metagenomics was to study the presence of viruses in seawater (Breitbart et al., 2002). The history of large high profile metagenomics projects continues in seawater in 2003 with Craig Venter's ambitious Global Ocean Sampling project. In this project, a sailing boat was equipped with a DNA sequencing lab and set off to sail across the globe to sample and sequence the DNA in seawater (Yooseph et al., 2007; Williamson et al., 2008). Early on, metagenomics was also used to investigate the microbial composition of an acid mine drainage system from which attempts to culture microbes had previously failed (Tyson et al., 2004; Hugenholtz, 2002). Early pioneering metagenomics projects thus gave researchers the first views of complex microbial ecosystems that had previously eluded them.

In the last ten years, many metagenomic sequencing projects have studied the human microbiome. The list of projects includes studies of the oral microbiota (Lazarevic et al., 2009), the diversity of the human intestinal flora (Eckburg et al., 2005), the human distal gut (Gill et al., 2006), and the prevalence of plasmids in the gut microbiome (Jones et al., 2010). Larger, recent metagenomic endeavors such as the Human Microbiome Project (Huttenhower et al., 2012) and the Metagenomics of the Human Intestinal Tract (Meta-HIT) (Qin et al., 2010) project have widely increased our understanding of the human microbiome and its connection to disease. Another large metagenomic study is The Earth Microbiome project whose aim was to construct a massive catalog of the genetic diversity of the entire planet (Gilbert et al., 2014).

2.2 Antibiotic resistance

Antibiotics are chemical or biological substances that either prevent the growth of, or kill, bacteria. After the introduction of modern antibiotics in the middle of the 20th century, the world has witnessed dramatic improvements to overall human health. Modern healthcare has grown highly reliant on effective antibiotics to treat and prevent bacterial infections (Rosenblatt-Farrell, 2009). However, due to overuse of antibiotics, bacteria are becoming resistant to many of our commonly used antibiotics (Neu, 1992; Andersson and Hughes, 2010, 2012; Arias et al., 2015). Because of the steady increase in the occurrence of antibiotic resistance, the World Health Organization (WHO) has recently classified the issue of drug resistance an important global problem and a threat to

human health (World Health Organization, 2015). Both the American Center for Disease Control (CDC) and the WHO advocate joint global responsibility and antibiotic stewardship programs to curb the spread of antibiotic resistance. Reducing antibiotic usage to maintain the efficacy of currently available treatments is particularly important considering that only two fundamentally new classes of antibiotics have been discovered since the 70's (Laxminarayan et al., 2013). There have been several examples of discoveries and development of antibiotic compounds where resistance emerged soon after the drug came into use (e.g. vancomycin and quinolone) (Levy and Marshall, 2004; Robicsek et al., 2006; Levine, 2006) The early days of rapid antibiotic discovery appear to have passed and the focus going forward must be to maintain the efficacies of antibiotics to try to avoid a future without efficacious antibiotics (Davies and Davies, 2010).

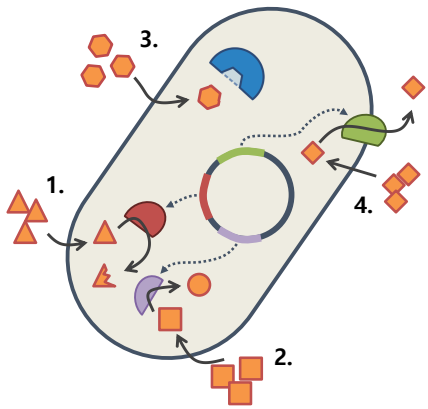


Figure 2.2: A bacterial cell displaying four examples of antibiotic resistance mechanisms: *1)* degradation of the antibiotic using an enzyme (red), *2)* modification of the antibiotic using an enzyme (purple), *3)* modification of the target enzyme (blue), and *4)* removal of the antibiotic using efflux pumps (green). Genes encoding proteins for the different mechanisms can be acquired via horizontal gene transfer, here symbolized by a circular plasmid with three antibiotic resistance genes encoding proteins performing these functions.

An organism is considered clinically resistant to an antibiotic if a drug that was once effective at treating an infection caused by that organism no longer is. Microbial organisms have a long history of withstanding antibiotics and typically have many methods at their disposal to help survive foreign threats (D'Costa et al., 2011; Davies and Davies, 2010; Sykes, 2010; Allen et al., 2010). Figure 2.2 illustrates four examples of common mechanisms for bacterial antibiotic resistance: *1*) degradation of the active substance, e.g. β-lactamases that break down β-lactam antibiotics, 2) modification of the antibiotic, e.g. aminoglyco-

side resistance genes that acetylate aminoglycocide molecules, *3*) alteration of the target site, e.g. fluoroquinolone resistance by mutations in target proteins such as gyrase, *4*) transporting the antibiotic out of the cell, e.g. via efflux pumps. Mechanisms such as these can be intrinsically encoded in the genome of an organism, but bacteria can also exchange genetic material with other bacteria (also with different species) in a process called horizontal gene transfer (Aminov and Mackie, 2007; Bennett, 2008). This process enables bacteria to pass genetic material between one another, including genes providing antibiotic resistance. Transfer of antibiotic resistance genes often occurs by transporting circular DNA constructs called plasmids between cells (Salyers et al., 2004). A single plasmid often contain multiple genes and may also contain mechanisms to efficiently spread these genes between cells (e.g., transposons, integrons, and conjugation systems) (Rajpara et al., 2009; Jacobsen et al., 2007; Hall and Collis, 1995).

Microorganisms have had thousands upon thousands of years to evolve effective ways to outwit their competition, and many species did so by developing the ability to produce compounds that negatively influenced the growth of their opponents (D'Costa et al., 2011). Consequently, microorganisms have, out of necessity, developed a large and diverse set of resistance mechanisms to outcompete their neighbors competing for the same niches. It has been hypothesized that microorganisms in both the environment and human microbiomes harbor a wide range of resistance mechanisms waiting to be discovered (Wright, 2010; Penders et al., 2013). Indeed, many commonly seen bacterial antibiotic resistance genes originate from the environment (Allen et al., 2010; Cantón, 2009). There is thus likely a flow of antibiotic resistance genes from harmless environmental or commensal bacteria into pathogenic species than can cause hard-to-treat infections. Because the underlying mechanisms of antibiotic resistance consist of distinct genetic components, it is possible to investigate their presence in a microbial community using DNA sequencing techniques. Metagenomics has been successfully applied to study the occurrence of novel antibiotic resistance genes in bacterial communities in the environment and in the human microbiome (Sommer et al., 2009; D'Costa et al., 2006; Boulund et al., 2012). Metagenomics has also been applied to examine the patterns of antibiotic resistance genes in polluted and pristine environments (Bengtsson-Palme et al., 2014; Kristiansson et al., 2011; D'Costa et al., 2011). Differences in the distributions of resistance genes in the human gut have also been identified connected with geographical origin (Forslund et al., 2013), and international travel (Bengtsson-Palme et al., 2015). A review by Schmieder and Edwards (2012) presents recent findings and some remaining challenges regarding the application of metagenomics to antibiotic resistance discovery.

2.3 Sequencing technologies

2.3.1 DNA sequencing

Deoxyribonucleic acid (DNA) is a biological molecule present in all living organisms that stores genetic instructions. DNA is replicated during the cell division process so that new cells inherit genetic traits from their parents. The molecule itself consists of basic building blocks called nucleotide bases that are connected together by a sugar backbone consisting of deoxyribose to form almost arbitrarily long strands. There are four types of nucleotides, normally abbreviated as single letters: adenine (A), cytosine (C), guanine (G), and thymine (T). The nucleotide bases can form pairwise bonds to one another; A with T, and C with G. This result is called base pairing and is crucial for the stability of DNA. The molecule is able to form helical (spiral) structures when two complementary strands of DNA base pair to one another (Watson and Crick, 1953), a property making them very stable.

An important concept in molecular biology is The Central Dogma (Crick, 1970) that states how information flows in biological systems. Information is stored in genes in DNA sequences that can be converted into proteins that perform biological functions. A gene can generally be considered a region of adjacent nucleotides that code for a protein. Within a protein coding gene, blocks of three consecutive nucleotides code for amino acids. There are 20 different standard amino acids that are normally coded for in DNA. Genes are translated into long chains of amino acids that generally fold into complex 3D structures to become proteins. Proteins are large biological molecules that perform many essential functions in cells. Because protein sequences are coded for in DNA sequences, the potential for a cell to produce a protein can be identified in that cell's DNA. This relation is essential for all the papers included in the thesis. For example, one of the basic assumptions underlying Papers I and III is that bacterial DNA sequences can be translated into amino acid sequences to identify protein sequences that can provide its host organism with resistance to antibiotic compounds.

To determine the DNA sequence of an organism, the DNA molecule(s) in that organism need to be extracted and the order of the nucleotide bases resolved. In 1975, a method now called Sanger sequencing was published (Sanger, 1977; Sanger and Coulson, 1975). Modern variants of this method are in some cases still in use because of its merits in sequencing relatively long continuous stretches of DNA (700-900 bases) at very high quality. The throughput of Sanger sequencing is limited, however, making it used mostly for smaller sequencing projects or when verifying specific genomic sequences. About

a decade ago the next-generation sequencing (NGS) era began when novel methods for high throughput DNA sequencing started becoming commercially available. These new technologies revolutionized sequencing and quickly replaced Sanger sequencing. One of the hallmarks of NGS technologies is that a large number of fragments are sequenced simultaneously, leading to the term "massively parallel sequencing". When sequencing with NGS technologies a method called shotgun sequencing is often used. In this method, DNA is extracted from samples of interest and enzymatically broken into many short pieces (i.e. like firing a shotgun at the DNA molecule). The fragmented DNA is put into a sequencing machine that determines the sequence of nucleotides in each fragment.

The sequence data returned from high throughput sequencing machines consists of reads, each representing a short section of DNA. Historically, NGS technologies produced reads as short as 25-50 nucleotides, but today's technologies normally generate significantly longer read lengths. The different NGS technologies can be grouped into categories. The technologies that are currently available in the market and most often encountered are based on techniques called sequencing by synthesis (e.g. Illumina, 454 pyrosequencing, and Ion Torrent), sequencing by ligation (e.g. ABI SOLiD), and single molecule sequencing (e.g. Pacific Biosciences SMRT and Nanopore). The read lengths of the different methods vary substantially, with SOLiD and Illumina on the lower end of the spectrum, producing reads between 75-300 bases in length. On the upper end, 454 pyrosequencing and Pacific Biosciences produce the longest reads at between 1000-20,000 nucleotides long, albeit often with general trade-offs in accuracy. A very good overview of the different NGS technologies is presented by Van Dijk et al. (2014).

Because of its low per-base cost, high throughput, and relatively low error rate, Illumina sequencing is routinely used in many sequencing projects and has become the most commonly employed technology in large-scale metagenomic sequencing studies. Sequencing using Illumina's technology works as follows (Mardis, 2008). First, DNA is extracted from a sample, then purified and fragmented into templates. The templates are attached to a solid surface on a flow cell where they are amplified to produce many copies of each template in small clusters on the surface. Each cluster thus consists of many copies of short identical DNA strands sitting close together. The template DNA strand must be amplified as the sequencing signal would be too faint to detect unless the clusters are made dense enough. The flow cell of an Illumina HiSeq 2500 machine can have up to 1 million clusters per square millimeter, which is one of the reasons for their immense throughput. The actual sequence of nucleotides in the strands represented by each cluster is determined by flowing a solution of free nucleotides across the flow cell. The nucleotides emit a small flash of

light when they bind to the next available position in the strands in the clusters. A high resolution camera detects the emitted light pulses, coded with different colors for each of the four nucleotides, and produces a sequence of images. From these images, the sequences of nucleotides in the strands in each cluster can be determined. This process is called base calling.

The DNA sequence data produced by NGS machines consists of short reads, which often requires special techniques for analysis. Unfortunately, sequencing machines are imperfect and occasionally return erroneous base calls (Treangen and Salzberg, 2012). The different technologies are prone to different types of errors. For example, insertion and deletion errors occasionally occur in 454 pyrosequencing and IonTorrent, i.e., a nucleotide that was not actually present in the real DNA strand is introduced into the sequence, or the converse when nucleotides are removed when they should actually be present. These errors are particularly common in regions with repeating nucleotides (e.g. multiple sequential A's). Illumina's technology, however, is more prone to randomly make incorrect base calls, leading to single nucleotide errors (substitutions) in the reads. The error rate of modern sequencers is typically in the range of $0.01-1\%$ (Huse et al., 2007; Dohm et al., 2008; Hansen et al., 2010; Robasky et al., 2013). Regardless of the type, errors introduce complexities in downstream analyses (Meacham et al., 2011). For example, errors make it more difficult to correctly find the position from where a read originates in a reference, or can severely complicate read assembly, making it important to consider and address the occurrence of errors in sequence data. Much research effort has been put into trying to model errors in reads and correct them, as highlighted by the vast numbers of available error correction algorithms, e.g. FASTX (Gordon and Hannon, 2010), HTQC (Yang et al., 2013), and many others (Ilie and Molnar, 2013; Meacham et al., 2011; Ilie et al., 2011; Liu et al., 2011; Yang et al., 2010; Kao et al., 2011).

2.3.2 Protein sequencing

Analogous to how DNA sequencing is used to determine the sequence of nucleotides in a DNA molecule, protein sequencing is used to determine the sequence of amino acids in a protein molecule. In the middle of the 20th century Edman (1950) presented a technique to determine the sequence of amino acids in proteins now called Edman degradation. Briefly, Edman degradation works by chemically marking and removing a single amino acid at a time from the N-terminal end of the protein chain. The process is repeated until the complete sequence is determined. After being automated in the late 1960s, the method became widely used. Currently, however, the method has been superseded by

either mass spectrometry-based methods, or by simply predicting the amino acid sequence directly from the nucleotide sequence (Coon, 2009).

Mass spectrometry analysis of a molecule begins an ionization step followed by measurement of the mass-to-charge ratio (m/z) using a mass analyzer. The analysis of biomolecules such as proteins became possible with the development of soft ionization techniques such as electrospray ionization (ESI) (Fenn et al., 1989) and Matrix-Assisted Laser Desorption/Ionization (MALDI) (Hillenkamp et al., 1991). In the ionization step, proteins or peptides (i.e. fragments of proteins) are converted into gaseous ions that enter the mass analyzer. Here, the charged ions are separated by their different m/z ratios in an electric or magnetic field. Examples of common mass analyzers in use today include various forms of ion traps, quadrupole (Q), Fourier-transform ion cyclotron resonance (FTICR), and time-of-flight (TOF) (Yates et al., 2009). The types vary in their underlying physical principles and performances. Seveeral variants are often combined to best utilize their different abilities, e.g. Q-TOF, using a quadrupole to separate the ions by mass, and then measuring their abundances in a TOF ion detector (Han et al., 2008). When sequencing proteins, two mass analyzers are normally used in sequence. In the first mass analyzer, the molecular ion of the intact protein or peptide is detected. The ion is then fragmented in a collision cell and further analyzed in the second mass analyzer. This process is called tandem MS. The output from MS analysis are spectra with peaks of the different m/z intensities. These spectra are compared with a precomputed database of known protein masses to determine what proteins were present in the original sample.

The term proteomics refers to the large-scale study of proteins, similar to how genomics relates to the study of genomes. Genomics can determine the presence of certain patterns in the DNA but is limited to predicting the potential for biological functions (genes are not always expressed). Proteomics has an advantage in that it allows for investigating the expressed protein performing actual functions in an organism. There are two main approaches to MS-based protein sequence characterization: bottom-up, and top-down (Meyer et al., 2011). In bottom-up proteomics, proteins are digested enzymatically (commonly using trypsin) into complex mixtures of peptides (McDonald and Yates 3rd, 2003; Banks et al., 2014). The bottom-up approach provides high sample throughput and is capable of identifying prokaryotic proteins with high accuracy (Yates et al., 2009). By contrast, top-down proteomics studies intact proteins instead of peptides (Chi et al., 2007; Macek et al., 2006; Swaney et al., 2010; Tran et al., 2011). Starting with intact proteins has the benefit of enabling the study of splice variants and post-translational modifications of proteins with high accuracy. However, the top-down approach has lower throughput and presents other additional complexities to produce the resulting spectra.

Two excellent reviews by Guerrera and Kleiner (2005) and Bantscheff et al. (2012) provide further details on how quantitative mass spectrometry is used in proteomics.

A common application of MS-based proteomics is the identification of microorganisms. The general concept of identification (and sometimes characterization of specific properties) is referred to as typing. Typing is used in varying contexts, including e.g. phenotyping (differentiation between organisms based on different phenotypes) and genotyping (classification by differences in genomes) (Welker and Moore, 2011). The classification of bacterial samples using fingerprints based on proteomic mass patterns MS has proven an efficient way of typing bacterial samples for many different species. In particular, MALDI-TOF MS has recently become part of the routine diagnostic workflow in many laboratories (Braga et al., 2012). MALDI-TOF MS is typically used for bacterial typing in clinical samples, i.e. to detect the presence of a select number of pathogenic bacterial species. There is also a potential for further developments of MS-based methods for clinical diagnostics. One example of a fairly recent development is the application of liquid chromatography tandem MS (LC MS/MS), utilizing a chromatographic column to provide better separation of the peptides prior to mass analysis, further improving the overall ability to accurately sequence individual peptides (Fournier et al., 2007; Issaq et al., 2005; McDonald and Yates 3rd, 2003). LC MS/MS has been shown capable of sub-species level identification for some bacterial species (Dworzanski et al., 2004, 2006; Jabbour et al., 2010a,b; Karlsson et al., 2012, 2015).

2.4 Challenges created by new technologies

Over the last 10-15 years we have continuously improved our ability to generate larger amounts of molecular data to address our biological questions and hypotheses. The developments in genomic and proteomic sequencing technologies combined with more powerful computing resources have shifted our focus from generating small specific data sets to answer specific questions, to generating massive amounts of data from which a wide range of different questions can be answered. For example, platforms such as Illumina's HiSeq 2500 can produce up one terabase of sequence data (10^{12} nucleotide bases), distributed across many billion reads, in a single run. Consequently, the amount of data in public repositories is accumulating at a rapid pace, already reaching petabase volumes (10^{15} bases) (Baker, 2010). The increase in sequence data submitted to the Sequence Read Archive (SRA) was recently reported to exceed the growth rate of hard drive storage capacity, and the European Nucleotide Archive (ENA) shows that the growth rate for submitted sequences

is exponential, doubling every 10 months (Figure 2.3 (Leinonen et al., 2011; Kodama et al., 2012; Cochrane et al., 2013).

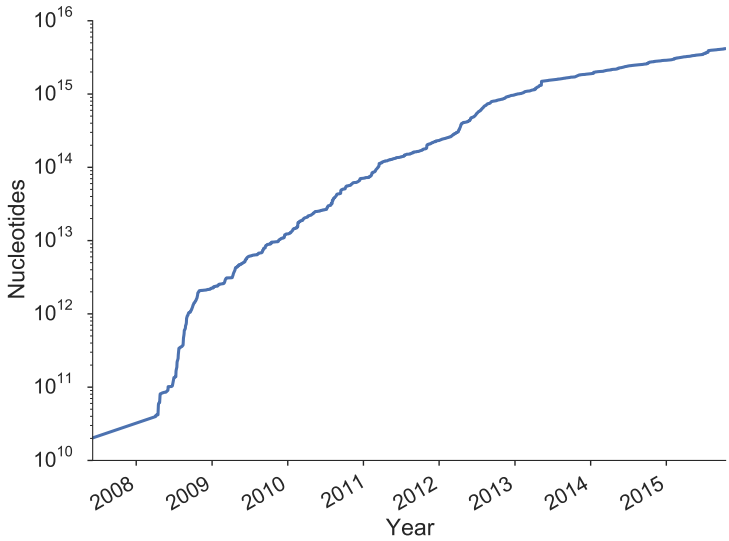


Figure 2.3: The Sequence Read Archive (SRA) is a repository for raw sequencing data that has exhibited exponential growth over the last years since the introduction of high throughput sequencing technology. Note the logarithmic axis for the number of nucleotides. Data from the National Center for Biotechnology Information (NCBI) SRA website (http://www.ncbi.nlm.nih.gov/Traces/sra/).

The rapidly increasing volumes of data that are accessible with new technologies have revolutionized the life sciences and how scientists can address questions and hypotheses in biology and medicine. Previously inaccessible biological information, such as the genetic basis of a bacterial community or all of the expressed proteins in a pathogen, can now be generated at low cost. However, the emergence of new high-throughput technologies has introduced new data analysis challenges, sometimes leaving established approaches in their wake. Most traditional methods for analyzing these types of data were not designed for the big, diverse, noisy, and heavily fragmented data generated by modern sequencing technologies. Applying traditional methods to data generated with these new technologies is therefore often infeasible. Thus, novel methods and approaches must be developed to unlock the full potential of this new generation of genomic and proteomic data.

3 Aims

The papers presented in this thesis were based on the following aims:

1. Develop and evaluate methodologies for the discovery of novel antibiotic resistance genes in fragmented metagenomic data.
2. Develop and implement a method for the quantification of genes in microbial communities using large volumes of fragmented metagenomic data.
3. Develop and evaluate the application of shotgun proteomics to bacterial identification in samples from complex bacterial mixtures.

The three aims relate to the included papers in the following way. The first aim addresses the ability to discover novel unknown antibiotic resistance gene sequences in short read data, the feasibility of which was relatively uncertain at that time. Papers I, II, and III describe how such an approach was designed and implemented, shown feasible, and had its predictions verified experimentally. The second aim addresses the difficulty in comparing very large metagenomic samples. Paper IV describes an implementation of a method that utilizes high performance computing cluster resources to efficiently perform gene quantification of very large metagenomic data sets. The third aim addresses the technical challenge of identifying the constituent members of complex bacterial mixtures in shotgun proteomics data. Paper V describes the general analysis workflow and presents the technological background of the method. In Paper VI we describe a method based on the concepts of Paper V, capable of sensitive identification of several clinically relevant pathogenic bacteria in mixed samples without a priori knowledge of the sample's bacterial makeup.

4 Methodological considerations

Most of the work in this thesis concerns the analysis of complex data sets consisting of fragmented biological sequences. This chapter will describe challenges related to the analysis of big data, presenting some approaches for how these challenges can be addressed. Further, this chapter is intended to complement the information on methodological concepts, data sets, and data formats that are presented in the papers. Refer to the methods section of each respective paper for complete details of the specific methods used in each of the studies.

4.1 Challenges with large data

New technologies for biological sequence generation can produce very large data sets, and the life sciences are facing numerous challenges in handling the great amounts of data being generated (Marx, 2013). The challenges associated with analyzing large data can be related to Moore's law. Originally formulated by Intel's co-founder Gordon E. Moore in 1965, Moore's law claims that the number of transistors in dense integrated circuits will double every two years (Schaller, 1997). Moore's statement has essentially held true since the beginning of the 1970s through to today, and has been predicted to hold for another decade. The continuous increase in computing power provides resources for analysis of larger and larger data sets. Mark Kryder made a similar observation regarding storage technologies in 2005. Kryder observed that the storage density of magnetic disks at that time was increasing at a pace much faster than the doubling time of processors put forth by Moore (Walter, 2005). The life sciences are currently experiencing very fast-paced development in regard to data generation capability, exhibiting a growth rate outpacing both Moore and Kryder's laws (Cochrane et al., 2013).

At a certain point, data sets become too large to analyze using normal methods and computers. The transition from regular data to "big data" varies extensively between fields, but some general properties can be identified. One definition is that data becomes big data when the size of the data sets pass the point where it is no longer feasible to apply regular computers and methods to store, manage, and analyze the data within reasonable time frames. The specific size at which this occurs shifts continuously as technology advances, but for sequence data in the life sciences this typically occurs as data sets approach terabase size. Big data is often described by characteristics called the three V's. The first definition of the three V's was laid forth in a research report in 2001 by META Group (Laney, 2001). In the report, the three V's of big data were defined as: Volume (the data must be large), Variety (integration of many different types of data), Velocity (analysis of data generated in real-time). For bioinformatics applications, volume is caused by the massive data generation capabilities of the latest sequencing technologies. Variety is caused by the combination and merging of multiple sources of data from e.g. genomics and proteomics and its associated metadata. Sequence data generation in e.g. DNA sequencing is currently at the tipping point where the velocity with which sequence data is generated is becoming a big factor, and the velocity component is likely to grow further in the coming years. Several real-time sequencing technologies are currently on their way to the market, which can be expected to produce a near continuous flow of sequence data to be analyzed, e.g. in routine sequencing of genomes and proteomes in clinical settings.

The alignment of biological sequences is central to many bioinformatics applications. Indeed, most of the methods presented in this thesis are based on sequence alignment. For example, the methods in Papers I and III use sequence alignment to find antibiotic resistance gene sequences in short DNA fragments. In Paper IV, the method is based on the alignment of from metagenomic data to reference sequences in order to determine the frequency of occurrence of the reference sequences in the metagenome. Paper VI uses sequencing alignment algorithms to identify bacterial species in shotgun proteomics data. The size of modern sequencing data sets is making timely alignment of sequences more challenging. In this context, traditional alignment algorithms, such as Smith-Waterman (Smith and Waterman, 1981) and Needleman-Wunsch (Needleman and Wunsch, 1970), are too inefficient because of their poor scaling ($\mathcal{O}(NM)$, in which N and M are the lengths of the two sequences being aligned). Research and development into sequence alignment has resulted in the traditional algorithms being superseded by more efficient algorithms for biological sequence alignment. The most notable being the now ubiquitous heuristic BLAST algorithm (Altschul et al., 1997) that revolutionized the entire field of bioinformatics, enabling fast and accurate sequence alignment. While still in widespread use, the performance of the BLAST algorithm is insufficient when it comes to align-

ing data at the scales of modern sequence data sets. Instead, novel algorithms such as Bowtie2 (Langmead and Salzberg, 2012), USEARCH (Edgar, 2010), BLAT (Kent, 2002), Vmatch (Kurtz, 2007), and BWA (Li and Durbin, 2009), have become instrumental for analysis of high throughput sequencing data. The key features of all of these algorithms are that they combine well-designed data structures, heuristics, and optimizations to provide efficient sequence alignment.

However, how to efficiently analyze data in the real world boils down to how the data is read and stored in the computer. The algorithms (and data structures that store the data) play a great part in how quickly and efficiently results can be generated from the raw data, with regards to computing power and memory consumption. Much research has focused on attaining high processing speed, and the most important part is often to strive for locality of reference, i.e. having the data as close to the processor as possible at all times, to minimize unnecessary transfer latencies. For big data analysis in life sciences, the limiting component tends to be input/output (I/O) performance, i.e. data access speeds; getting the data from storage to CPU (Trelles et al., 2011). Thus, the challenge often lies in feeding the data to the analysis algorithm as fast as possible. For example, the time required to read 1 terabyte of data from a solid-state drive (SSD) is almost 40 minutes (at 500 MiB/s), and from a regular hard drive it will take over 3 hours (at 100 MiB/s).

Compression algorithms are widely used to make data smaller, e.g. to minimize the storage consumption of large files. Compression can also be used as a way to increase the amount of information transferred from network and disk per time unit. Using compression can thus help reduce the limitations of I/O throughput by sacrificing some CPU power to minimize the total number of bytes required to transfer from the storage media. Several general purpose compression algorithms often perform well on DNA and protein sequence data because these data are often stored in ASCII-based formats that contain redundant information (e.g. similar sample headers or duplicate sequences). Implementations of general purpose compression algorithms such as gzip and bzip2 are typically available by default in most operating systems. Because of their ubiquity, we implemented support for sequence data compressed with gzip in the method described in Paper IV to increase the net data throughput when transferring data over the network. There are also specialized compression algorithms developed specifically for use with sequencing data in FASTA and FASTQ formats, e.g. MFCompress (Pinho and Pratas, 2013) and DSRC2 (Roguski and Deorowicz, 2014). These methods typically achieve higher compression ratios than general purpose compression algorithms, and some parallel implementations (e.g. the aforementioned DSRC2) provide compression and decompression speeds of more than 500 MiB/s. Thus, using

specialized and highly performant compression algorithms can help alleviate the challenges related to big data I/O when working with sequence data, which will improve the ability to generate insight and results from large biological data.

4.2 Distributed computing

One approach to analyzing big data is to parallelize the analysis, e.g. by splitting the data into smaller pieces and analyzing each of them concurrently. Parallelization, i.e. the concept of performing analysis of several subtasks in parallel, can substantially reduce the total time for analysis. Most computers today have processors capable of limited parallelism, but this is generally not enough to handle the large amounts of data commonly encountered in modern sequence data sets. Distributed computing resources such as high performance computing (HPC) clusters and cloud computing is often required to achieve the necessary data throughput to analyze big data. Using distributed computing systems to run analyses in parallel enables the study of larger data sets than what was previously possible, opening up new possibilities e.g. for inferring correlations in large data sets.

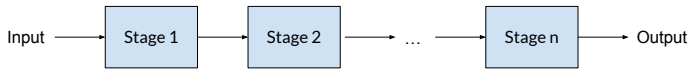
In order to effectively utilize distributed computing resources, algorithms must, in most situations, be modified from their single processor implementations. To reap the full benefits of parallelization, this often requires extensive knowledge of the algorithm and the hardware the application is expected to run on. There are two general approaches to parallelization: data parallelism and task parallelism (Subhlok et al., 1993). Data parallelism divides a data set into smaller pieces, operating on smaller parts of the data to reduce the overall computational time required. To get the largest benefit from data parallelism there must not be any dependencies between different parts of the data. If the problem cannot be cleanly separated into parts without requiring communication between the separate parts, the problem typically parallelizes less efficiently. Task parallelism divides problems on a higher level, decomposing problems into separate tasks that can often be operated completely independently of each other, allowing for efficient parallelization. The time to perform an analysis will only decrease if concurrency can be achieved.

If paralleization of a program is going be beneficial can be analyzed using Amdahl's law. Amdahl's law states that the maximum speedup that can be achieved for a program of which P is the proportion that can be parallelized, using N processors, is $S(N) = \frac{1}{(1-P)+\frac{P}{N}}$ (Rodgers, 1985). In situations where only a part of a sequential application can be parallelized, the parameter P

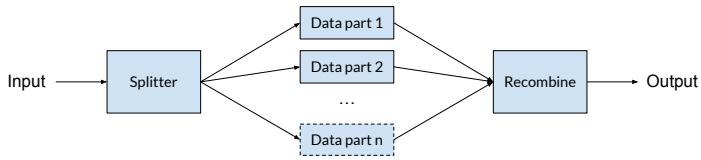
is substantially less than 1. Even if the part can be parallelized to run almost instantly using several CPUs, the entire application would still have its total runtime limited by the non-parallelizable part. For perfectly parallel problems, i.e. problems that can be split into parts where there is complete independence between the different parts, the parameter P is almost equal to 1. The total runtime of such problems can, in theory, be reduced by a factor of N using N processors. Examples of perfectly parallel problems in bioinformatics include e.g. the alignment of DNA sequence reads to reference sequences, which in theory can be parallelized almost perfectly by splitting up the read data into smaller parts (i.e. data parallelism).

Several patterns reoccur when problems are parallelized. These patterns can be referred to as algorithmic skeletons (Cole, 1991; Mattson et al., 2004). Abstracting parallelism into skeletons allows for reusable implementations and simplifies management of distributed computations. Here, three skeletons that are widely applicable to bioinformatics computing problems are briefly described and illustrated (Figure 4.1). A commonly encountered skeleton is the pipe skeleton (Figure 4.1a), ubiquitous in unix-like environments. The pipe skeleton fits well into the unix ecosystem, perhaps as a direct result of the unix philosophy that each program should do one thing and do it well, and programs should be composable via a unified text-stream interface (McIlroy et al., 1978). In the pipe skeleton, several components are chained together to form long pipelines of separate components that act in several stages. This allows parallelization by running each of the components on separate processing units. Drawbacks include that all pipeline components depend on one another serially, making the pattern less suited for workflows involving speed-limiting stages, such as heavily I/O-bound operations. The map skeleton is another common skeleton (Figure 4.1b), which essentially consists of *split, execute, merge* operations condensed into a single method. The data is split into multiple parts, each of which execute the same task on different parts of the data in parallel. As each part finishes, the results from each part of the data are combined into a single result. The well-known MapReduce algorithm (Dean and Ghemawat, 2008) was designed based on this skeleton. The MapReduce algorithm has been used in several bioinformatics applications (e.g. sequence alignment Schatz (2009), Nguyen et al. (2011), and distributed data transformation Schumacher et al. (2014)). Finally, the farm skeleton, sometimes also referred to as the master-worker skeleton, has worker processes that execute tasks in parallel. The tasks are distributed and managed by the emitter/master process (Figure 4.1c). The farm skeleton allows for the execution of dissimilar subtasks in ad-hoc systems, and is therefore sometimes also referred to as the task farm skeleton. It bears superficial similarity to the map skeleton, but the workers are generally under more direct control by the emitter/master process and the tasks can be different, unlike the map skeleton. For more detail about

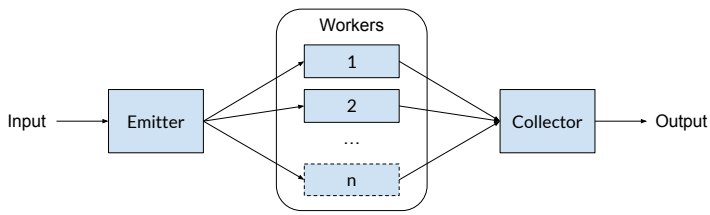
the concepts behind parallelization using algorithmic skeletons the reader is invited to read the review by González-Vélez and Leyton (2010).



(a) The pipe skeleton. This skeleton performs different subtasks in separate stages in a contiguous stream, in which the output of each subtask is connected to the input of the next. A pipeline can e.g. be used to efficiently filter data using several successive criteria.



(b) The map skeleton. This skeleton implements a function that splits the data into n parts that are executed in n separate processes. The results from each part are combined in a final stage to produce the complete result. This can e.g. be used to efficiently process convert an array of floating point numbers to integers in parallel.



(c) The farm skeleton. This skeleton has a program often called the master process that splits the input into n jobs that are sent (emitted) to n separate worker processes. The workers can be located on completely different computers, connected via a network. The master process receives the results from each worker process and combines the results to form the complete result. This skeleton can e.g. be used to efficiently perform read mapping on computer clusters.

Figure 4.1: Three examples of algorithmic skeletons, i.e. skeleton for parallelization. Data is input from the left and the processed output exits to the right.

The problems described in this thesis are mostly of the data parallel type and thus amenable to parallelization. For example, in Papers I and III the method is parallelizable by splitting the metagenomes into pieces of suitable size and the analysis could be run in parallel on distributed computing systems e.g. using

the map skeleton. The method implemented in Paper IV uses several skeletons simultaneously to parallelize on different levels. On the first level, the raw data is split into pieces that are sent to worker processes in a distributed computing system similar to the farm skeleton. On the second level, each worker process performs analysis of its set of data using the pipeline skeleton. The method for bacterial proteotyping presented Papers V and VI combines the pipeline and farm skeletons to analyze several samples in parallel in its workflow. In general, most sequence alignment-based methods are well-suited for parallelization. This essentially means that it should be possible to decompose most alignment-based methods into task or data parallel implementations that parallelize perfectly, fully utilizing all available computing resources in a parallel computing system (disregarding any distribution overhead). Readers are invited to read more about computational solutions management and analysis of large-scale data in the review by Schadt et al. (2010).

4.2.1 Computer hardware for bioinformatics

Regular workstation or laptop computers are unsuited for handling large data. Bioinformatics researchers generally utilizes high powered computing servers with high performance processors (often at least two quad- or octa-core processors per machine), large amounts of fast RAM (generally several hundred gigabytes), and large hard drive arrays for storing raw data, temporary intermediate data, and results. The analysis of large data sets sometimes requires more computing resources than single computers can typically provide. In such cases, high performance computing (HPC) clusters can be used. For researchers in Sweden, the Swedish National Infrastructure for Computing (SNIC) (http://www.snic.vr.se/) provides HPC resources located at several universities across Sweden. These resources are hosted and managed by local centers at each site. In this thesis, the vast majority of the analyses were done on the HPC systems at Chalmers, managed by the Chalmers Centre for Computational Science and Engineering (C3SE) (http://c3se.chalmers.se).

The main operating systems used within bioinformatics are based on GNU/Linux. There are many different distributions of Linux developed and maintained especially for scientific use, such as Scientific Linux (https://www.scientificlinux.org/) sponsored by Fermi National Accelerator Laboratory and CERN, and BioLinux (http://environmentalomics.org/bio-linuxx/) (Field et al., 2006) sponsored by the Natural Environment Research Council (NERC). These specialized science-oriented Linux distributions are generally based on established Linux distributions, such as Ubuntu Linux (BioLinux) or Red Hat Enterprise Linux/CentOS (Scientific Linux). Nearly all methods presented in this thesis (Papers I, III, IV,

and VI) were designed to run primarily on the CentOS-based Linux systems on the local HPC infrastructure provided by C3SE at Chalmers.

An alternative to using local HPC installations is to use cloud computing resources. Cloud computing is becoming more common as a recent developments in the last five to ten years have provided viable, cost effective, and flexible analysis platforms. Vendors such as Amazon, Google, and Microsoft, offer cloud computing resources for which an increasing number of bioinformatics methods are developed (Stein, 2010). The flexibility they afford also makes them useful as backends for web-based bioinformatics tools. For example, the well-known metagenomics analysis platform MG-RAST (Meyer et al., 2008) uses cloud computing resources to analyze uploaded data. The method presented in Paper IV was constructed with the long-term goal to include the ability to utilize cloud computing resources to analyze large metagenomes. Cloud service provides in general offer a different type of billing model compared to regular HPC installations, making them well suited for the varying and intermittent workloads commonly seen in bioinformatics research. Publishing large data sets so that they are available for analysis via cloud computing resources has been discussed as a way to increase the accessibility of such data sets, making it easy for researchers to analyze the data without having to download the raw data (Schatz et al., 2010).

4.3 Data used in this thesis

Data from several different public data sets and reference databases are used in this thesis. This section describes a selection of the data sets and reference databases employed. Each section also contains a brief explanation of the most often encountered formats used to store the different types of data.

4.3.1 Metagenomics data

Several data sets of metagenomic data from the human microbiome were used in this thesis. The American National Institute of Health's Human Microbiome Project (The NIH HMP Working Group, 2009) is the largest and most varied and deepest sequence human associated metagenomic data set available today. The metagenomic data set that was published in 2012 comprises 3.5 terabases of raw reads sampled at up to 18 body sites (e.g. oral, skin, gut, etc.) from 242 individuals, produced using the Illumina platform with 101 base pair long reads (Methé et al., 2012; Huttenhower et al., 2014). Another large data

set of gut microbiome shotgun metagenomes was produced as part of the Metagenomics of the Human Intestinal Tract (Meta-HIT) project (Qin et al., 2010), which was generated from faecal samples of 124 European individuals. The published data consists of raw reads produced with the Illumina platform using 45-75 base pair long reads. When the data set was published in 2010 it was the largest single public metagenomic data set of its kind with its 576.7 gigabases of sequence data. A third shotgun metagenomic data set with gut microbe metagenomes was published by Qin et al. (2012), produced for a type 2 diabetes study on 345 Chinese individuals. The data set consists of 378.6 gigabases of raw read data and was also produced on the Illumina platform (150 base pair reads). The metagenomic data from human microbiomes were used to search for novel resistance genes in Paper I and III, and to evaluate and optimize the method for distributed gene quantification in Paper IV.

Environmental metagenomes from several projects were also used in this thesis. The CAMERA database (Seshadri et al., 2007), containing data from the Global Ocean Sampling (GOS) expedition (Venter et al., 2004; Yooseph et al., 2007), along with several other smaller environmental metagenomic data sets, was also screened for *qnr* genes in Paper I. Several additional smaller metagenomic data sets were also used in Papers I and III, most prominently metagenomic data from polluted environments in Patancheru, India (Bengtsson-Palme et al., 2014).

The FASTQ format has become the de facto standard format for storing sequence data produced by next generation DNA sequencing machines (Cock et al., 2010). FASTQ stores the sequenced reads along with information on the quality of each nucleotide base in an ASCII-based format, making it easy manually inspect and modify using any text editor. Briefly, each record in a FASTQ file consists of at least four lines (Figure 4.2).

The quality values in FASTQ files correspond to an integer mapping of the probability, p, that the corresponding base was called incorrectly. The quality values, commonly denoted Phred scores (Ewing et al., 2005), were first introduced with Sanger sequencing and is encoded as $Q = -10\ log_{10}p$. For example, a 99.99% probability results in a Phred score of 40. The range of possible scores is not limited and can in theory extend upwards as technology improves. The range of quality scores e.g. from Illumina's sequencing machines is however normally expected to lie within the range 0-41, and is represented as the printable ASCII characters `!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ`.

```
@FBSEQ-READ30:1:v3:1337:5150#151204 length=30
AGTCGAATCGCATGCTCAGTCAGCATGCAG
+FBSEQ-READ30:1:v3:1337:5150#151204 length=30
,5;?BHIIIIIIIIIIIIHIHI@51/-+)
```

Figure 4.2: Example of a single record for a 30 nucleotide long DNA read in FASTQ format. The first line contains the record identifier and always starts with an `@` character. The second line contains the raw sequence letters in ASCII format (i.e. ATGC and N for unknown). The third line begins with a `+` character and is optionally followed by the same record identifier. The fourth line contains the base calling quality values encoded in printable ASCII; this line is always the same length as the sequence line. In this example, the identifier line is repeated for the quality scores, but this is rarely encountered in real sequencing data because it is redundant and space consuming. The format specification allows the sequence and score lines to be of any length, and optionally allows them to also be wrapped at any length as well, a feature that is also rarely encountered in FASTQ files today.

4.3.2 Proteomics data

Proteomic data sets in this thesis were all produced as part of an EU-project, The Tailored Treatment (http://www.tailored-treatment.eu/), which aims to develop personalized diagnostics of bacterial and viral infections. The proteomics data produced were generated specifically for the project by tandem MS/MS using a Thermo Fischer Scientific Q Exactive™ Hybrid Quadrupole-Orbitrap mass spectrometer. Samples were prepared using a proprietary flow cell technology, LPI™ Flowcell (Nanoxis Consulting), used to immobilize bacterial cells for more controlled protein digestion and extraction. Shotgun proteomics data generated from pure bacterial cultures and mixtures of pure cultures were used in Paper V and VI. The mass spectra, initially stored in Thermo Fischer Scientific's proprietary raw data format, were run through X!!Tandem (Bjornson et al., 2008) to generate peptide sequences. Samples generally consisted of 1500-3000 usable spectra, producing peptides in the 6-45 amino acid length range. Please refer to Papers V and VI for more details on the specifics of the proteomics data used.

Vendors of MS equipment typically have their own proprietary raw data formats in which data is normally stored in opaque binary files. These binary files can often be converted to open text-based standard formats such as e.g. mzXML using vendor supplied libraries. The mzData format was the first attempt at an open standard for MS data, but vendors wanted to wait for the standard to be finalized before implementing support. In the meantime, a new

format called mzXML was developed to fill the need for an open standard that the scientific community could use Pedrioli et al. (2004). However, the scientific community identified some shortcomings in the mzXML format and have, under the organization of the Human Proteome Organization's Proteomics Standards Initative (HUPO-PSI), developed the mzML standard as its successor. The mzML format was first presented by HUPO-PSI in 2008 (Deutsch, 2008), aiming to combine the best elements of the earlier two formats, and the scientific community is converging towards exclusively using this format. The XML-based standards are well defined, making parsing them with any standard XML-parser straightforward, and most open source proteomics software are capable of reading and writing these formats.

4.3.3 Reference databases

The National Center for Biotechnology Information (NCBI) supplies a wide range of publicly available databases. The GenBank database (Benson et al., 1999) is a public database of nucleotide and peptide sequences, submitted from researchers from all over the world. GenBank was used in all of the papers in this thesis. It effectively acts as a central repository for all publicly available nucleotide and protein sequences. However, since practically any sequence is accepted, the quality of the stored sequences and their annotation can vary (Harris, 2003; Nilsson et al., 2008). The most recent release (210.0, released 2015-Oct-15) contains $202,237,081,559$ bases, from $188,372,017$ reported sequences (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt). In addition to the general sequence repository, the NCBI also supplies an open access database of annotated, high quality, reference genome sequences for several organisms (e.g., viruses, bacteria, and eukaryotes), called the Reference Sequence database (RefSeq) (Pruitt et al., 2007). The sequences for each organisms are stored as single records for the complete biological molecules, and RefSeq currently contains more than $9,000$ bacterial genomes. Another important database maintained by NCBI that is used throughout this thesis is the taxonomy database, NCBI Taxonomy (Federhen, 2012). Although NCBI Taxonomy is considered a non-authoritative source for taxonomic information, the taxonomy database represents essentially the entire tree of life with convenient links to other NCBI resources such as RefSeq sequences and annotations. Papers I and III screened GenBank and RefSeq for the presence of *qnr* antibiotic resistance genes. Paper VI used the NCBI Taxonomy and RefSeq genomes to identify bacterial presence in samples.

5 Summary of results

This chapter provides summaries of the results of the six papers included in the thesis. The overall aims, findings, and conclusions from each paper are presented to further make it easier to understand each paper's respective contribution to the research field.

5.1 Papers I, II, III

The first three papers are closely connected. In Paper I we developed and applied a method to metagenomic data sets to discover novel antibiotic resistance genes. The novel genes discovered in the first paper were then verified experimentally in Paper II, by expressing the discovered genes in a bacterial host to confirm their antibiotic resistance properties. Paper III further improved the method developed in Paper I and applied it to much larger data sets consisting mainly of short read metagenomes.

In Paper I, *A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences*, we developed and implemented a method enabling the identification of a certain type of antibiotic resistance genes called *qnr* in metagenomes consisting of short DNA fragments. The *qnr* genes were discovered approximately 15 years ago, and provide a mechanism of resistance with the potential to rapidly spread between bacteria using horizontal gene transfer (Martínez-Martínez et al., 1998). Together, the *qnr* genes form a family of plasmid mediated resistance genes that provide a wide range of different types of bacteria with low to moderate levels of resistance to commonly used antibiotics in the fluoroquinolone family (e.g. ciprofloxacin). Similar to most antibiotic resistance genes, *qnr* genes are assumed to originate from environmental bacteria (Poirel and Rodriguez-Martinez, 2005). Over the last decade several families of *qnr* genes have been discovered and characterized, but their

true prevalence and diversity remain unclear. In particular, environmental and host-associated bacterial communities have been hypothesized to maintain a large and unknown collection of *qnr* genes that could be mobilized into pathogens.

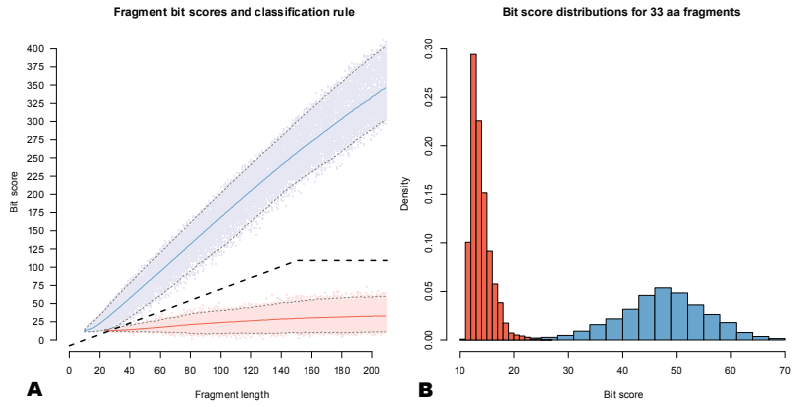


Figure 5.1: A) Empirical bit scores of fragmented *qnr* (blue) and non-*qnr* (red) peptide sequences at varying fragment lengths. Notice the clear distinction between *qnr* and non-*qnr* sequences, separated by the *qnr* classifier function (dashed line). B) Bit scores distributions for 33 amino acids long *qnr* and non-*qnr* fragments, showing the separation in bit score distribution that makes it possible to discern *qnr* from non-*qnr* at 33 amino acid long fragments (equivalent to 100 base pairs long nucleotide fragments).

The *qnr* proteins belong to a group called pentapeptide repeat proteins. Proteins in this group display a specific repeating pattern in their amino acid sequence, consisting of repeating subunits of five amino acid residues. Because of this specific pattern, sequences from *qnr* proteins can be accurately detected using hidden Markov models (HMMs). The strength of HMMs is that they are able to capture the repeating pattern in the amino acid sequence while still allowing high variability in regions of the sequence where there is little conservation across different variants of *qnr*. Furthermore, HMMs are also computationally efficient (Eddy, 2011) and the vast amounts of data generated by shotgun metagenomics can therefore be used to explore the diversity of *qnr* genes in more detail. The main finding of Paper I was that the *qnr* family of antibiotic resistance genes is possible to identify using HMMs. We developed an HMM based on all known *qnr* gene variants that, in combination with a custom classifier that merges fragment length information with expected bit scores against the HMM, achieves high performance even for fragment lengths down to 100 base pairs (Figure 5.1A). The model was able to accurately identify the amino acid sequence of *qnr* fragments and distinguish them from other similar pentapeptide repeat sequences that were not derived from *qnr* genes

(Figure 5.1B). In our study we also discovered several fragments that indicate a presence of previously unknown *qnr* gene variants in the environment. The method described in this paper significantly improves the sensitivity and specificity of identification and annotation of qnr genes in nucleotide sequence data. The predicted novel putative qnr genes in the metagenomic data support the hypothesis of a large and uncharacterized diversity within this family of resistance genes in environmental bacterial communities. Our implementation of the method was made in Python 2.7 for use in Linux environments and is freely available at http://bioinformatics.math.chalmers.se/qnr/. Paper I was published in *BMC Genomics*, **13**:695 (Boulund et al., 2012).

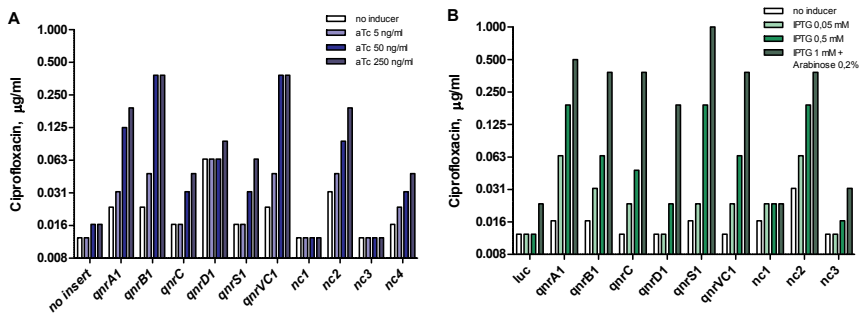


Figure 5.2: Minimum inhibitory concentration (MIC) of known and novel *qnr* gene variants when expressed and overexpressed in *E. coli* using IPTG and arabinose.

The second paper, *Functional verification of computationally predicted qnr genes*, is a follow-up study of the results from Paper I. The paper describes a bacterial expression platform for evaluating the resistance phenotype of antibiotic resistance genes, in which *Escherichia coli* was chosen as the host organism. The expression platform was evaluated using synthesized genes of several types of well-known *qnr* genes and novel *qnr* candidate genes discovered in Paper I. By using inducible recombinant expressions systems the functionality of four identified *qnr* candidates were evaluated. Expression and overexpression of several known *qnr* genes as well the novel candidates provided fluoroquinolone resistance that increased with elevated inducer concentrations. One of the main results of Paper II is that two of the putative *qnr* genes discovered in Paper I provide antibiotic resistance when expressed in an *E. coli* host (Figure 5.2). Papers I and II also serve as an important example underscoring how computational methods can be used in exploratory studies to generate novel hypotheses and intermediary results that can be verified in the lab. The combination of a computational model and recombinant expression systems provides opportunities to explore and identify novel antibiotic resistance genes

in both genomic and metagenomic datasets. Paper II was published in *Annals of Clinical Microbiology and Antimicrobials*, **12**:34 (Flach et al., 2013).

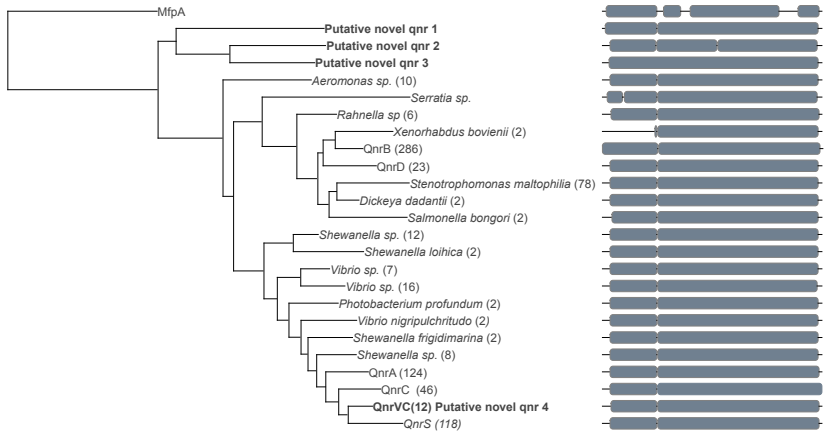


Figure 5.3: Gene tree showing plasmid-mediated and chromosomal *qnr* gene families detected and discovered in Paper III. The tree is rooted to MfpA, a pentapeptide protein similar to *qnr*. Novel sequences discovered in Paper III are depicted in bold. To the right of the tree is a simplified representation of a multiple alignment of all sequences.

Paper III revisited and optimized the method developed in Paper I for larger and more fragmented data sets. We applied the improved method to an almost ten times larger data set, comprising more than 5 terabases of sequence data. In total, 256,520 potential *qnr* gene fragments were identified, from which 669 putative *qnr* genes were reconstructed. These gene sequences included all previously described plasmid-mediated *qnr* gene families. Twenty-one of the 699 identified *qnr* genes were reconstructed from metagenomes, of which four were novel and previously undescribed. Three of these novel putative genes were only distantly related to known qnr gene families, whereas the fourth shared 73% sequence identity to *qnrVC6* (Figure 5.3). The *qnr* gene predictions presented in this study provide the basis for follow-up experiments, similar to those in Paper II, to validate the fluoroquinolone resistance phenotypes of the identified gene sequences.

5.2 Paper IV

Paper IV, titled *Tentacle: distributed quantification of genes in metagenomes*, addressed the challenge of working with very large scale metagenomic data sets. Using shotgun metagenomics, microbial communities can be sequenced

with high resolution, generating data sets containing billions of DNA fragments. With the continuous development of sequencing technologies, the size of metagenomic data sets are expected to increase. The trend of increasing data set sizes shows no signs of slowing down.

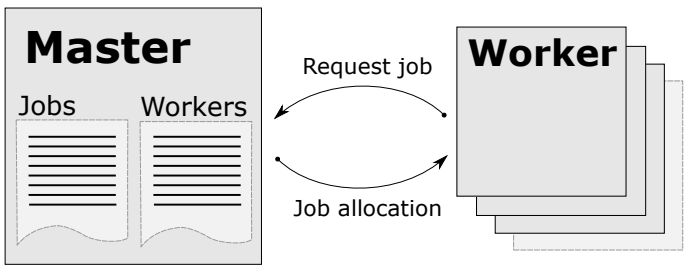


Figure 5.4: Schematic view of Tentacle's master-worker interaction and responsibilities. The master process is manages the list of jobs and maintains a list of currently available and engaged worker processes. Each worker process registers with the master process and start receiving jobs when it comes online. As worker processes complete their allocated jobs they request new jobs until the list of jobs is depleted.

An application where the analysis of very large amounts of metagenomic data is useful is for the assessment of differences between bacterial communities (Jones et al., 2010; Tringe et al., 2005). This could for example be comparisons of the gut microbiome between sick and healthy individuals. To perform such comparisons it is necessary to quantify the presence of genes or other genomic features of interest (Delmont et al., 2013). To perform gene quantification requires that each read in the metagenomic sample is compared to a reference (e.g. databases of known genes or bacterial genomes). After all reads have been compared the number of matches to the references are counted and can be used as a basis for statistical comparative analysis (Kristiansson et al., 2009). Sequence comparison (i.e. sequence alignment) is a task of high computational complexity, both requiring well designed algorithms and powerful computer hardware to perform efficiently at the scale required for large modern metagenomic data sets. Novel methods that can efficiently process the growing volumes of sequence data are necessary for the accurate analysis and interpretation of existing and upcoming metagenomes. The aim of this project was to enable researchers to perform gene quantification in metagenomic data sets in sizes of up to several terabases (10^{12} nucleotide bases).

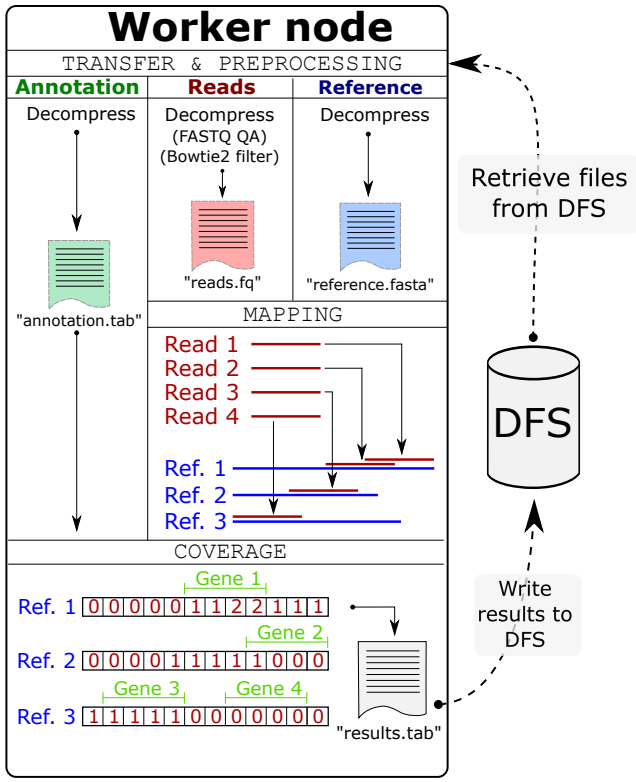


Figure 5.5: Processing that take place in Tentacle workers. Workers receive job instructions from the master process, and then autonomously retrieves the required files from the distributed file system (DFS) without involvement of the master process. The workers decompress and preprocesses annotations, reads, and reference sequences in contiguous in-memory streams, avoiding unnecessary disk operations. When reads have been mapped to the reference sequences, the coverage across annotated regions in the references is computed and written directly back to the DFS, without involvement of the master process. The worker then requests a new job from the master process.

In Paper IV we developed Tentacle, a framework that allows researchers to use distributed computational resources (e.g. high-performance computing clusters) for gene quantification in large metagenomes. We designed Tentacle to employ a dynamic master-worker approach in which worker nodes can be started independently and join the computations whenever (Figure 5.4). Each worker node streams all data via the network and processes everything within an in-memory stream (i.e. like a unix pipeline, see section 4.2). How worker nodes process data is depicted in Figure 5.5. The master-worker approach that

Tentacle is designed upon is akin to the farm skeleton previously described in section 4.2. The fact that data in metagenomes is inherently independent is central to the model of distributing computations in Tentacle. Because each read is independent of other reads, the metagenomic data can be split and processed in parallel without requiring any inter-worker process communication. Tentacle was tested on a high performance computing (HPC) cluster to evaluate how the method scales with increasing data sizes and number of worker nodes. The results indicate that our dynamic master-worker approach scales almost linearly with increasing computing resources (Figure 5.6), making it well suited for analyzing very large metagenomic data sets. We also designed Tentacle to be as modular and extensible as possible, to make the framework as flexible as possible. Because of the extensibility, Tentacle already supports six commonly used sequence aligners, giving researchers the possibility to use Tentacle to analyze their metagenomes without needing to learn the intricacies of a different mapping algorithm than what they are used to. We wanted to make it easy to adapt Tentacle to different applications in metagenomics and easy to integrate into existing workflows.

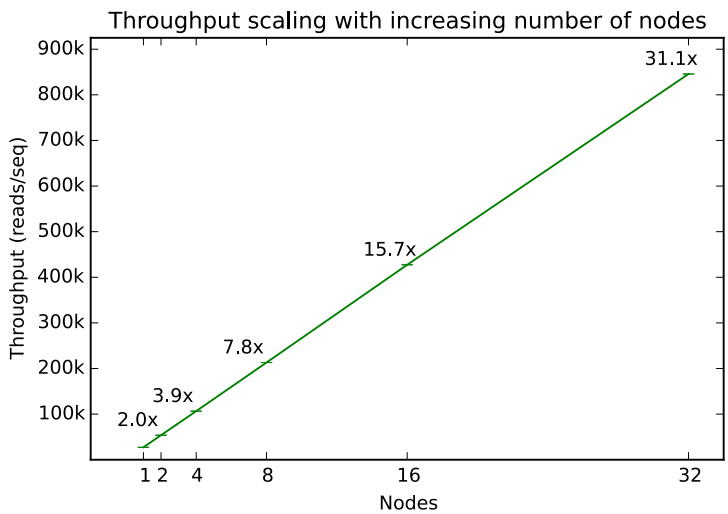


Figure 5.6: Tentacle scales almost perfectly linearly with increasing number of nodes. In this example, Tentacle was run on a subset of data from a metagenomic data set, split into equally sized parts. The number of parts analyzed by Tentacle was equal to the number of utilized HPC nodes, i.e. making Tentacle analyze twice the amount of data for each doubling of utilized nodes. For each doubling of the number of computing nodes, the total throughput of Tentacle effectively doubles, but the wall-clock time remains constant.

The fundamental result of Paper IV is a new method for distributed gene quantification in very large metagenomic data sets. The method, called Tentacle, is based on a bioinformatics pipeline consisting of components that work together to perform gene quantification in metagenomic data, which is in turn distributed across several nodes in high-performance computer clusters in a data parallel manner. This exploits the independence between reads to allow parallel read mapping and gene quantification that scales very well with increasing computing resources (Figure 5.6). We designed Tentacle to run on Linux systems using Python 2.7, and it is published as open source under the GNU General Public License (v3). Source code and documentation is freely available at http://bioinformatics.math.chalmers.se/tentacle/. Paper IV was published in *GigaScience*, **4**:40 (Boulund et al., 2015).

5.3 Papers V, VI

Methods for rapid and reliable microbial identification are essential in modern health care. The ability to detect and correctly identify pathogenic species is, for example, necessary for accurate diagnosis of infectious diseases and effective treatment. MS-based shotgun proteomics is a technology that can rapidly characterize large parts of the expressed genes of microorganisms present in a sample.

Paper V is a self-standing book chapter that gives a thorough background on MS and MS-based bacterial identification, and introduces the concept of proteotyping and the bioinformatics workflow required for the proteotyping analysis (Figure 5.7).

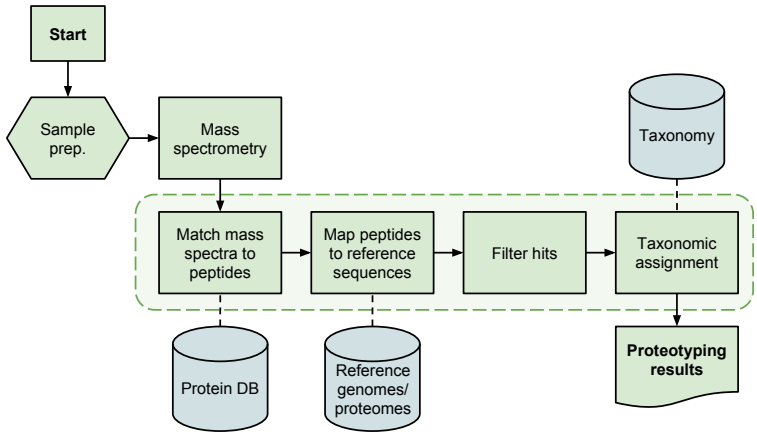


Figure 5.7: An overview of the proteotyping workflow presented in Paper V. The workflow begins in the upper left corner with a sample begin generated by bottom-up tandem MS. Using an extensive protein database, the mass spectra are converted to peptides. Each peptide is then mapped to a set of curated reference genomes/proteomes, allowing multiple hits. The hits to the reference genomes are filtered to remove noise and spurious hits. The remaining hits are positioned in a taxonomic tree and the lowest common ancestor algorithm is applied to determine which peptides are suited for inferring what bacterial species were present in the original sample.

In Paper VI we developed a new computational method for proteotyping based on the principles outlined in Paper V. The method uses data from shotgun MS for detecting bacteria and determining their taxonomic affiliations. The MS data is highly fragmented, consisting of many small protein fragments (i.e. peptides). Each peptide is compared against a large database consisting of full genome sequences. A single peptide can match to several genome sequences, e.g. if the peptide comes from a protein common to many different bacteria. After determining what peptide matches to which genomes, the method applies the lowest common ancestor (LCA) algorithm. To be certain that a specific bacterium was present in the sample the method needs to find peptides that can uniquely identify that bacterium, i.e. a peptide that matches only a specific genome. The LCA algorithm makes it possible to automatically identify peptides that can be used to determine what organisms are likely to be present in a sample.

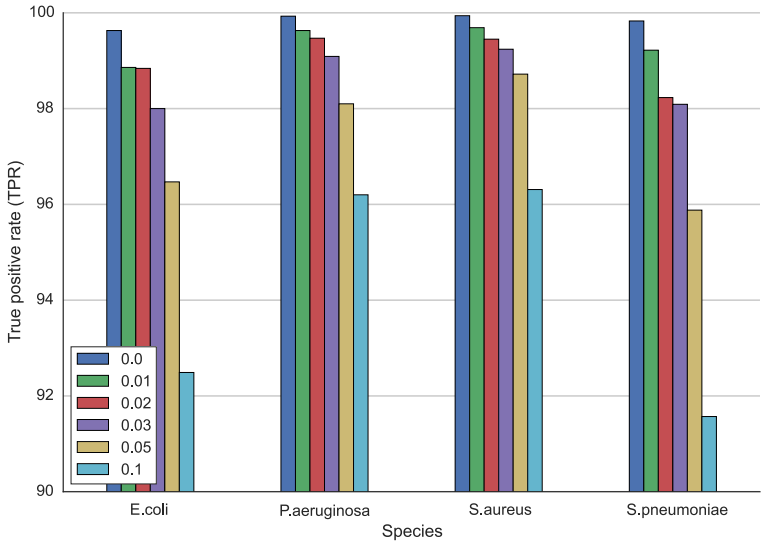


Figure 5.8: Proteotyping performance on pure in silico samples of single bacterial species. The true positive rate (TPR) shows what percentage of peptides could be assigned to the correct species in each of the pure culture samples. The effect of random mutations is shown by the decrease in TPR as mutation rate increases up to 10%. *E. coli* and *S. pneumoniae* exhibit the most prominent decrease in TPR, whereas the identification of the two other species is more robust to single point mutations.

We evaluated the method using four commonly encountered clinically-relevant bacterial species, *Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Streptococcus pneumoniae*. The data used in the evaluation was from both simulated data generated by in silico peptide digestion, and data from pure culture samples, generated by tandem mass spectrometry (LC-MS/MS). The results indicated that the method was able to correctly classify peptides at a true positive rate between 91.2% to 98.8% for the different species (Figure 5.8). Furthermore, the method also performed well for the identification of the individual species in mixed samples. By correcting the estimated species abundances with the information on species-specific variation in the proportion of identified peptides, accurate detection of the relative abundances of the respective species was possible (Figure 5.9)

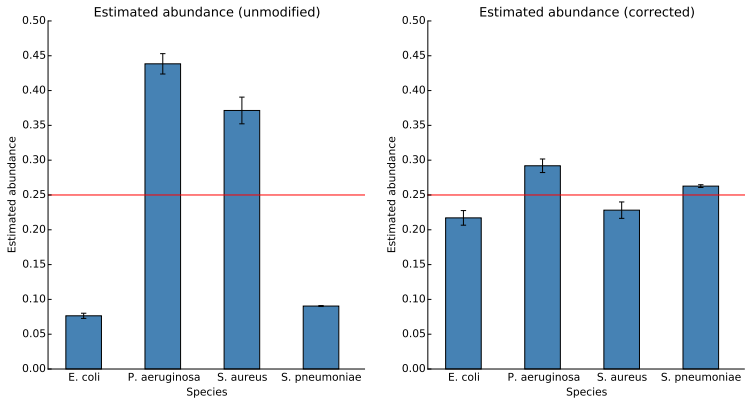


Figure 5.9: Proteotyping performance on sample of mixed bacterial cultures in 1:1:1:1 ratio. The left figure shows the raw estimated abundances directly using the identified number of peptides belonging to each species. The right figure shows the abundance estimations of same sample after applying a correction that adjusts the observed counts by a factor determined by results from samples of pure cultures. The horizontal line represents the true relative abundance of each bacterial species.

The main results of papers V and VI are that they show that bottom-up MS proteomics can be used to not only accurately detect single species of bacteria in pure cultures, but also perform identification in complex mixtures more akin to real clinical samples. The methods presented in the papers are proof-of-concepts that demonstrate the potential of these techniques as a rapid tool for diagnostics of infectious diseases.

6 Conclusions and future perspectives

Biology and life sciences are currently experiencing unprecedented developments in data generation capability. There is a need for novel methods to help solve the challenges posed by these new and vast data sets. This thesis shows examples of novel approaches and techniques that can be used to handle big data in this area of research. The papers presented in this thesis describe several methods aimed at facing the challenges that arise when analyzing such large data sets.

In Paper I and III, we developed new methods to screen large data sets of fragmented DNA sequence data for the presence of novel, previously uncharacterized *qnr* antibiotic resistance genes. In Paper II, the proposed novel sequences were validated to provide decreased susceptibility to fluoroquinolone antibiotics when expressed in *E. coli*. These results show that it is possible to make accurate predictions of novel antibiotic resistance genes directly from short read data without prior assembly. The results also support the hypothesis that is an unknown diversity of *qnr* genes in environmental bacterial communities. The methodologies presented in these papers are suitable for studying even larger metagenomes. By capitalizing on the inherent data parallelism that is possible in the of analysis of raw reads, the method should be amenable to parallelization in a scalable manner, which means that the method is applicable to future larger metagenomic data sets. Finally, the methods described in Paper I and III are general and can be applied to other classes of resistance genes. Thus, these methods can therefore be important tools for characterizing the large and still unknown diversity of resistance genes believed to exist in bacterial communities.

In Paper IV we developed Tentacle, a framework for parallel, distributed quantification of genetic elements, capable of handling the continuously growing

metagenomic data sets. The master-worker approach we implemented was shown to scale very well with increasing resources, making applicable for the quantification and analysis of metagenomes at the terabase size range and above. Tentacle thereby removed many of the big data challenges associated with the analysis of modern metagenomic data sets. Tentacle demonstrates that parallelization and distribution can improve performance on generally I/O bound analysis tasks such as read alignment. Because of its modular implementation, the underlying distribution framework core in Tentacle is, in theory, possible to apply to other tasks than metagenomic gene quantification, e.g., genome resequencing or RNA sequencing. Furthermore, we plan to improve the accessibility of Tentacle by increasing user friendliness, adding support more distribution schedulers, adding support for the utilization of cloud computing resources, and implementing new features using the Tentacle framework, such as improved gene quantification via high resolution binning.

In Paper V we developed the concept of proteotyping, i.e. using shotgun proteomics for identification and characterization of microbial samples. Based on this concept, we produced an implementation in Paper VI that was shown to be able to accurately identify bacterial species in both single cultures and mixtures of common pathogenic bacteria. The method shows promise for future applications in clinical diagnostics of infectious diseases. In particular, proteotyping has the potential to be completely culture independent and thereby reduce the time to diagnosis substantially. Proteotyping also shows promise in being able to accurately identify multispecies infections and for combining microbial identification with additional characterization. Work has already started in this direction with the aim to develop an integrated diagnostics platform capable of identifying and characterizing microbial organisms by e.g. determining their virulence and antibiotic resistance properties.

In conclusion, biology and medicine are experiencing a big data revolution that brings new opportunities for discovery and learning. The rapidly increasing volumes of diverse and heavily fragmented data present new challenges with regards to data analysis. New approaches and methods for data analysis, such as those presented in this thesis, are required to allow researchers to dive into previously inaccessible depths of information, to reach deeper understandings of the inner workings of biological systems.

Bibliography

Allen, H. K., Donato, J., Wang, H. H., Cloud-Hansen, K. A., Davies, J., and Handelsman, J. (2010). Call of the wild: antibiotic resistance genes in natural environments. *Nature reviews. Microbiology*, 8(4):251–9.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Aminov, R. I. and Mackie, R. I. (2007). Evolution and ecology of antibiotic resistance genes. *FEMS microbiology letters*, 271(2):147–61.

Andersson, D. I. and Hughes, D. (2010). Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature reviews. Microbiology*, 8(4):260–71.

Andersson, D. I. and Hughes, D. (2012). Evolution of antibiotic resistance at non-lethal drug concentrations. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, 15(3):162–72.

Arias, C. A., Murray, B. E., Ph, D., and Murray, B. E. (2015). A new antibiotic and the evolution of resistance. *New England Journal of Medicine*, 372(12):1168–1170.

Baker, M. (2010). Next-generation sequencing: adjusting to data overload. *Nature Methods*, 7(7):495–499.

Banks, C. A. S., Lakshminarasimhan, M., and Washburn, M. P. (2014). *Shotgun Proteomics*, volume 1156. Springer New York.

Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, 404(4):939–965.

Bengtsson-Palme, J., Angelin, M., Huss, M., Kjellqvist, S., Kristiansson, E., Palmgren, H., Larsson, D. J., and Johansson, A. (2015). The human gut

microbiome as a transporter of antibiotic resistance genes between continents. *Antimicrobial Agents and Chemotherapy*, (August):00933–15.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., and Larsson, D. G. J. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Frontiers in Microbiology*, 5(December):1–14.

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, 153 Suppl(January):347–57.

Benson, D. a., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. a., and Wheeler, D. L. (1999). GenBank. *Nucleic acids research*, 27(1):12–7.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., and Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of human biology*, 40(October 2015):463–71.

Bjornson, R. D., Carriero, N. J., Colangelo, C., Shifman, M., Cheung, K. H., Miller, P. L., and Williams, K. (2008). X!!Tandem, an improved method for running X!Tandem in parallel on collections of commodity computers. *Journal of Proteome Research*, 7(1):293–299.

Boulund, F., Johnning, A., Pereira, M. B., Larsson, D. G. J., and Kristiansson, E. (2012). A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. *BMC genomics*, 13(1):695.

Boulund, F., Sjögren, A., and Kristiansson, E. (2015). Tentacle: distributed quantification of genes in metagenomes. *GigaScience*, 4(1):40.

Braga, P. A. C., Tata, A., Santos, V. G., Barreiro, J. R., Schwab, N. V., Santos, M. V., Eberlin, M. N., and Ferreira, C. (2012). Bacterial identification: from the agar plate to the mass spectrometer. *RSC Advances*, pages 994–1008.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14250–5.

Cantón, R. (2009). Antibiotic resistance genes from the environment: a perspective through newly identified antibiotic resistance mechanisms in the clinical setting. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 15 Suppl 1:20–5.

Chi, A., Bai, D. L., Geer, L. Y., Shabanowitz, J., and Hunt, D. F. (2007). Analysis of intact proteins on a chromatographic time scale by electron transfer dissociation tandem mass spectrometry. *International journal of mass spectrometry*, 259(1):197–203.

Cochrane, G., Alako, B., Amid, C., Bower, L., Cerdeño-Tárraga, A., Cleland, I., Gibson, R., Goodgame, N., Jang, M., Kay, S., Leinonen, R., Lin, X., Lopez, R., McWilliam, H., Oisel, A., Pakseresht, N., Pallreddy, S., Park, Y., Plaister, S., Radhakrishnan, R., Rivière, S., Rossello, M., Senf, A., Silvester, N., Smirnov, D., Ten Hoopen, P., Toribio, A., Vaughan, D., and Zalunin, V. (2013). Facing growth in the european nucleotide archive. *Nucleic Acids Research*, 41(Database issue):30–5.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771.

Cole, M. I. (1991). Algorithmic skeletons : Structured management of parallel computation. *Computing*, page 170.

Coon, J. J. (2009). Collisions or electrons? protein sequence analysis in the 21st century. *Analytical chemistry*, 81(9):3208–15.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227:561–563.

Davies, J. and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews : MMBR*, 74(3):417–33.

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

Dean, J. and Ghemawat, S. (2008). Mapreduce : Simplified data processing on large clusters. *Communications of the ACM*, 51(1):1–13.

Delmont, T. O., Simonet, P., and Vogel, T. M. (2013). Mastering methodological pitfalls for surviving the metagenomic jungle. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 35(8):744–54.

Deutsch, E. (2008). mzML: A single, unifying data format for mass spectrometer output. *Proteomics*, 8:2776–2777.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105.

Dworzanski, J., Snyder, A., and Zhang, H. (2004). Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. *Analytical Chemistry*, 76(8):2355–2366.

Dworzanski, J. P., Deshpande, S. V., Chen, R., Jabbour, R. E., Snyder, a. P., Wick, C. H., and Li, L. (2006). Mass spectrometry-based proteomics combined with bioinformatic tools for bacterial classification. *Journal of Proteome Research*, 5(1):76–87.

D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., and Wright, G. D. (2011). Antibiotic resistance is ancient. *Nature*, 477(7365):457–461.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., and Relman, D. a. (2005). Diversity of the human intestinal microbial flora. *Science (New York, N.Y.)*, 308(5728):1635–8.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, 7(10):e1002195.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics (Oxford, England)*, 26(19):2460–1.

Edman, P. (1950). Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica*, 4:283–293.

Ewing, B., Ewing, B., Hillier, L., Hillier, L., Wendl, M. C., Wendl, M. C., Green, P., and Green, P. (2005). Base-calling of automated sequencer traces using. *Genome Research*, (206):175–185.

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):136–143.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71.

Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N., and Thurston, M. (2006). Open software for biologists: from famine to feast. *Nature biotechnology*, 24(7):801–803.

Flach, C.-F., Boulund, F., Kristiansson, E., and Larsson, D. J. (2013). Functional verification of computationally predicted qnr genes. *Annals of clinical microbiology and antimicrobials*, 12(1):34.

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, a., and Bork, P. (2013). Country-specific antibiotic use practices impact the human gut resistome. *Genome Res*, 23(7):1163–1169.

Fournier, M. L., Gilmore, J. M., Martin-brown, S. A., and Washburn, M. P. (2007). Multidimensional separations-based shotgun proteomics. (816):3654–3686.

Gilbert, J. a., Jansson, J. K., and Knight, R. (2014). The Earth microbiome project: successes and aspirations. *BMC Biology*, 12(1):69.

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

González-Vélez, H. and Leyton, M. (2010). A survey of algorithmic skeleton frameworks: High-level structured parallel programming enablers. *Software - Practice and Experience*, 40(12):1135–1160.

Gordon, A. and Hannon, G. J. (2010). Fastx-toolkit. FASTQ/A short reads pre-processing tools (unpublished).

Guerrera, I. C. and Kleiner, O. (2005). Application of mass spectrometry in proteomics. *Bioscience Reports*, 25:71–93.

Hall, R. M. and Collis, C. M. (1995). Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Molecular microbiology*, 15(4):593–600.

Han, X., Aslanian, A., and Yates, J. R. (2008). Mass spectrometry for proteomics. *Current opinion in chemical biology*, 12(5):483–90.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):245–9.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131.

Harris, D. J. (2003). Can you bank on genbank? *Trends in Ecology & Evolution*, 18(7):315–317.

Hillenkamp, F., Karas, M., Beavis, R. C., and Chait, B. T. (1991). Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193A–1203A.

Hooper, L. V. and Gordon, I. J. (2001). Commensal host-bacterial relationships in the gut. *Science*, 292(5519):1115–1118.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome biology*, 3(2):REVIEWS0003.

Hugenholtz, P., Goebel, B., and Pace, N. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, 180(18):4765–4774.

Huse, S. M., Huber, J. a., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel dna pyrosequencing. *Genome biology*, 8(7):R143.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. a., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. a., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G. a., Buhay, C. J., Busam, D. a., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S. G., Chen, I.-M. a., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. a., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Michael Dunne, W., Scott Durkin, a., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. a., Forney, L. J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B. J., Hamilton, H. a., Harris, E. L., Hepburn, T. a., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. a., Keitel, W. a., Kelley, S. T., Kells, C., King, N. B., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C. a., Dwayne Lunsford, R., Madden, T., Mahurkar, A. a., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavromatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. a., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Pop, M., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J. a., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata,

N., Segre, J. a., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. a., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. a., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. a., Wellington, C., Wetterstrand, K. a., White, J. R., Wilczek-Boney, K., Wu, Y., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. a., Highlander, S. K., Methé, B. a., Nelson, K. E., Petrosino, J. F., Weinstock, G. M., Wilson, R. K., and White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Huttenhower, C., Knight, R., Brown, C. T., Caporaso, J. G., Clemente, J. C., Gevers, D., Franzosa, E. a., Kelley, S. T., Knights, D., Ley, R. E., Mahurkar, A., Ravel, J., and White, O. (2014). Advancing the microbiome research community. *Cell*, 159(2):227–30.

Ilie, L., Fazayeli, F., and Ilie, S. (2011). HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 27(3):295–302.

Ilie, L. and Molnar, M. (2013). RACER: Rapid and accurate correction of errors in reads. *Bioinformatics (Oxford, England)*, pages 1–4.

Issaq, H. J., Chan, K. C., Janini, G. M., Conrads, T. P., and Veenstra, T. D. (2005). Multidimensional separation of peptides for effective proteomic analysis. *Journal of Chromatography B*, 817(1):35–47.

Jabbour, R. E., Deshpande, S. V., Wade, M. M., Stanford, M. F., Wick, C. H., Zulich, A. W., Skowronski, E. W., and Peter Snyder, A. (2010a). Double-blind characterization of non-genome-sequenced bacteria by mass spectrometry-based proteomics. *Applied and Environmental Microbiology*, 76(11):3637–3644.

Jabbour, R. E., Wade, M. M., Deshpande, S. V., Stanford, F., Wick, C. H., Zulich, A. W., and Snyder, a. P. (2010b). Identification of yersinia pestis and escherichia coli strains by whole cell and outer membrane protein extracts with mass spectrometry-based proteomics identification of yersinia pestis and escherichia coli strains by whole cell and outer membrane protei. *Journal of Proteome Research*, pages 3647–3655.

Jacobsen, L., Wilcks, A., Hammer, K., Huys, G., Gevers, D., and Andersen, S. R. (2007). Horizontal transfer of tet(m) and erm(b) resistance plasmids from food strains of lactobacillus plantarum to enterococcus faecalis jh2-2 in the gastrointestinal tract of gnotobiotic rats. *FEMS microbiology ecology*, 59(1):158–66.

Jefferson, K. (2004). What drives bacteria to produce a biofilm? *FEMS Microbiology Letters*, 236(2):163–173.

Jones, B. V., Sun, F., and Marchesi, J. R. (2010). Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. *BMC genomics*, 11:46.

Kao, W.-C., Chan, A. H., and Song, Y. S. (2011). Echo: a reference-free short-read error correction algorithm. *Genome research*, 21(7):1181–92.

Karlsson, R., Davidson, M., Svensson-Stadler, L., Karlsson, A., Olesen, K., Carlsohn, E., and Moore, E. R. B. (2012). Strain-level typing and identification of bacteria using mass spectrometry-based proteomics. *Journal of Proteome Research*, 11(5):2710–2720.

Karlsson, R., Gonzales-Siles, L., Boulund, F., Svensson-Stadler, L., Skovbjerg, S., Karlsson, A., Davidson, M., Hulth, S., Kristiansson, E., and Moore, E. R. (2015). Proteotyping: Proteomic characterization, classification and identification of microorganisms – a prospectus. *Systematic and Applied Microbiology*, 38(4):246–257.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664.

Kodama, Y., Shumway, M., and Leinonen, R. (2012). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1):10–13.

Kreft, J.-U. (2004). Biofilms promote altruism. *Microbiology*, 150(8):2751–2760.

Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., Weijdegård, B., Söderström, H., and Larsson, D. G. J. (2011). Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PloS one*, 6(2):e17038.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics (Oxford, England)*, 25(20):2737–8.

Kurtz, S. (2007). The Vmatch large scale sequence analysis software features of vmatch.

Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety. *Application Delivery Strategies*, 949(February 2001):4.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Laxminarayan, R., Duse, A., Wattal, C., Zaidi, A. K. M., Wertheim, H. F. L., Sumpradit, N., Vlieghe, E., Hara, G. L., Gould, I. M., Goossens, H., Greko, C., So, A. D., Bigdeli, M., Tomson, G., Woodhouse, W., Ombaka, E., and Peralta, A. Q. (2013). Antibiotic resistance — the need for global solutions. *The Lancet infectious diseases*, 3099(13):1057–1098.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osterås, M., Schrenzel, J., and François, P. (2009). Metagenomic study of the oral microbiota by illumina high-throughput sequencing. *Journal of microbiological methods*, 79(3):266–71.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1).

Levine, D. P. (2006). Vancomycin: a history. *Clinical Infectious Diseases*, 42(Supplement 1):S5–S12.

Levy, S. B. and Marshall, B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. *Nature medicine*, 10(12 Suppl):122–9.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

Liu, Y., Schmidt, B., and Maskell, D. L. (2011). DecGPU: distributed error correction on massively parallel graphics processing units using CUDA and MPI. *BMC bioinformatics*, 12:85.

Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006). Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, 5(5):949–958.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9:387–402.

Martínez-Martínez, L., Pascual, A., and Jacoby, G. A. (1998). Quinolone resistance from a transferable plasmid. *Lancet*, 351(9105):797–9.

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498(7453):255–260.

Mattson, T. G., Sanders, B. A., and Massingill, B. L. (2004). *Patterns for parallel programming*. Pearson Education.

McDonald, W. H. and Yates 3rd, J. R. (2003). Shotgun proteomics: integrating technologies to answer biological questions. *Current opinion in molecular therapeutics*, 5(3):302–309.

McIlroy, M. D., Pinson, E. N., and Tague, B. A. (1978). Unix time-sharing system: Foreword. *Bell System Technical Journal*, 57(6):1899–1904.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics*, 12(1):451.

Methé, B. a., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., Gevers, D., Petrosino, J. F., Abubucker, S., Badger, J. H., Chinwalla, A. T., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Hallsworth-Pepin, K., Lobos, E. a., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. a., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V. R., Brooks, P., Buck, G. a., Buhay, C. J., Busam, D. a., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S., Chen, I.-M. a., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. a., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Michael Dunne, W., Scott Durkin, a., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. a., Forney, L., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Haas, B. J., Hamilton, H. a., Harris, E. L., Hepburn, T. a., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. a., Keitel, W. a., Kelley, S. T., Kells, C., Kinder-Haake, S., King, N. B., Knight, R., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C. a., Dwayne Lunsford, R., Madden, T., Mahurkar, A. a., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavrommatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. a., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J. a., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. a., Shannon, W. D.,

Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. a., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. a., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. a., Wellington, C., Wetterstrand, K. a., White, J. R., Wilczek-Boney, K., Qing Wu, Y., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. a., Highlander, S. K., Weinstock, G. M., Wilson, R. K., and White, O. (2012). A framework for human microbiome research. *Nature*, 486(7402):215–221.

Meyer, B., Papasotiriou, D. G., and Karas, M. (2011). 100% protein sequence coverage: a modern form of surrealism in proteomics. *Amino acids*, 41(2):291–310.

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

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–53.

Neu, H. (1992). The crisis in antibiotic resistance. *Science (New York, N.Y.)*, 257(7):837–42.

Nguyen, T., Shi, W., and Ruden, D. (2011). CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. *BMC research notes*, 4(1):171.

Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N., and Larsson, K.-H. (2008). Intraspecific its variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary bioinformatics online*, 4:193–201.

Pedrioli, P. G. a., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, 22(11):1459–1466.

Penders, J., Stobberingh, E. E., Savelkoul, P. H. M., and Wolffs, P. F. G. (2013). The human microbiome as a reservoir of antimicrobial resistance. *Frontiers in microbiology*, 4(April):87.

Pinho, A. J. and Pratas, D. (2013). Mfcompress : a compression tool for fasta and multi-fasta data. pages 3–5.

Poirel, L. and Rodriguez-Martinez, J. (2005). Origin of plasmid-mediated quinolone resistance determinant qnra. *Antimicrobial agents*, 49(8).

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1):61–65.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S. S. S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-m., Hansen, T., Le, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Jian, M., Zhou, Y., Li, Y., Zhang, X., Guarner, F., Qin, N., Yang, H., Wang, J. J., Brunak, S., Dore, J., Le Paslier, D., Doré, J., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., and Ehrlich, S. D. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

Rajpara, N., Patel, A., Tiwari, N., Bahuguna, J., Antony, A., Choudhury, I., Ghosh, A., Jain, R., Ghosh, A., and Bhardwaj, A. K. (2009). Mechanism of drug resistance in a clinical isolate of vibrio fluvialis: involvement of multiple plasmids and integrons. *International journal of antimicrobial agents*, 34(3):220–5.

Robasky, K., Lewis, N. E., and Church, G. M. (2013). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56–62.

Robicsek, A., Jacoby, G. a., and Hooper, D. C. (2006). The worldwide emergence of plasmid-mediated quinolone resistance. *The Lancet infectious diseases*, 6(10):629–40.

Rodgers, D. P. (1985). Improvements in multiprocessor system design. *ACM SIGARCH Computer Architecture News*, 13(3):225–231.

Roguski, L. and Deorowicz, S. (2014). DSRC 2—industry-oriented compression of FASTQ files. *Bioinformatics*, 30(15):2213–2215.

Rosenblatt-Farrell, N. (2009). The landscape of antibiotic resistance. *Environmental health perspectives*, 117(6):245.

Salyers, A. a., Gupta, A., and Wang, Y. (2004). Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends in microbiology*, 12(9):412–6.

Sanger, F. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Sanger, F. and Coulson, A. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441–448.

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature reviews. Genetics*, 11(9):647–57.

Schaller, R. (1997). Moore's law: past, present and future. *Spectrum, IEEE*, 34(6):52–59.

Schatz, M., Langmead, B., and Salzberg, S. (2010). Cloud computing and the DNA data race. *Nature biotechnology*, 28(7):691–693.

Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics (Oxford, England)*, 25(11):1363–9.

Schloss, P. D. and Handelsman, J. (2004). Status of the microbial census. *Microbiology and molecular biology reviews : MMBR*, 68(4):686–691.

Schmieder, R. and Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches. *Future Microbiology*, 7(1):73–89.

Schumacher, A., Pireddu, L., Niemenmaa, M., Kallio, A., Korpelainen, E., Zanetti, G., and Heljanko, K. (2014). SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. *Bioinformatics (Oxford, England)*, 30(1):119–20.

Seshadri, R., Kravitz, S. a., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics. *PLoS biology*, 5(3):e75.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147:195–197.

Sommer, M. O. A., Dantas, G., and Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science (New York, N.Y.)*, 325(5944):1128–31.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome biology*, 11(5):207.

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of Bacteriology*, 194(16):4151–4160.

Subhlok, J., Stichnoth, J. M., O'hallaron, D. R., and Gross, T. (1993). Exploiting task and data parallelism on a multicomputer. In *ACM SIGPLAN Notices*, volume 28, pages 13–22. ACM.

Swaney, D. L., Wenger, C. D., and Coon, J. J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of proteome research*, 9(3):1323–1329.

Sykes, R. (2010). The 2009 Garrod lecture: the evolution of antimicrobial resistance: a Darwinian perspective. *The Journal of antimicrobial chemotherapy*, 65(9):1842–52.

Taylor, L. H., Latham, S. M., and Woolhouse, M. E. (2001). Risk factors for human disease emergence. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 356(1411):983–989.

The NIH HMP Working Group (2009). The NIH Human Microbiome Project. *Genome Research*, 19(12):2317–2323.

Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E., Catherman, A. D., Durbin, K. R., Tipton, J. D., Vellaichamy, A., Kellie, J. F., and Li, M. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, 480(7376):254–258.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46.

Trelles, O., Prins, P., Snir, M., and Jansen, R. C. (2011). Big data, but are we ready? *Nature reviews. Genetics*, 12(3):224.

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

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):1–9.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, 304(5667):66–74.

Walter, C. (2005). Kryder's law. *Scientific American*, 293(2):32–3.

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Welker, M. and Moore, E. R. B. (2011). Applications of whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry in systematic microbiology. *Systematic and Applied Microbiology*, 34(1):2–11.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583.

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M., and Venter, J. C. (2008). The Sorcerer II Global Ocean Sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PloS one*, 3(1):e1456.

World Health Organization (2015). Worldwide country situation analysis: response to antimicrobial resistance. Technical Report April.

Wright, G. D. (2010). Antibiotic resistance in the environment: a link to the clinic? *Current opinion in microbiology*, 13(5):589–94.

Yang, X., Dorman, K. S., and Aluru, S. (2010). Reptile: representative tiling for short read error correction. *Bioinformatics (Oxford, England)*, 26(20):2526–33.

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., Zhao, F., and Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC bioinformatics*, 14(1):33.

Yates, J. R., Ruse, C. I., and Nakorchevsky, A. (2009). Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering*, 11:49–79.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. a., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H., Mashiyama, S. T., Joachimiak, M. P., van Belle, C., Chandonia, J.-M., Soergel, D. a., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A., Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M., and Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS biology*, 5(3):e16.