THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Analysis of large-scale metagenomic data

Fredrik Boulund

CHALMERS | GÖTEBORGS UNIVERSITET

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2013

Analysis of large-scale metagenomic data
Fredrik Boulund
NO 2013:17
ISSN 1652-9715

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

Typeset with LaTeX.
Printed in Göteborg, Sweden 2013

Analysis of large-scale metagenomic data

Fredrik Boulund

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

Abstract

The topic of this thesis is the analysis of large data sets of DNA sequence data produced from modern high-throughput DNA sequencing machines. Using such machines to sequence the genetic content of a microbial community produces a metagenome. This thesis comprises three research papers, all connected to the study of large metagenomic data sets.

In the first paper, we developed a method for discovering fragments of fluoroquinolone antibiotic resistance genes in short fragments of DNA. The method uses hidden Markov models for identifying *qnr* genes in short DNA fragments. Cross-validation showed that our method for classifying short fragments has high statistical power even for fragments as short as 100 base pairs, a length commonly encountered in modern next-generation sequencing data.

In the second paper, the putative *qnr* genes identified in the first paper were verified using wet-lab experiments. This was a follow-up study to validate the findings from the first paper. An expression system for *qnr* genes in *Escherichia coli* hosts was developed and used to evaluate the resistance phenotype of the novel gene candidates discovered in the first paper.

In the third paper, we developed an easy-to-use high performance method for distributed gene quantification in metagenomic sequence data. It leverages high-performance computing resources to provide high throughput while maintaining sensitivity. This enables efficient and accurate gene quantification, suitable for use in comparative metagenomics.

Next-generation DNA sequencing has had a big impact on molecular biology. As the size of the produced data sets increases, there is an equally increasing need for methods suited for the analysis of such data sets. This thesis presents several new methods that are well adapted to analysis of modern terabase-sized metagenomic data sets.

Keywords: metagenomics, DNA analysis, big data, antibiotic resistance, hidden Markov models, high-performance computing, distributed computing

Acknowledgements

First off, I would like to wholeheartedly thank my advisors; Erik Kristiansson, Anders Sjögren, Marija Cvijovic. You are all great and inspiring! Next up are all the other wonderful people at the department that make it such a joy to come to work in the mornings, some of which deserve to have their names put in here: Viktor Jonsson, Anna Larsson, Mariana Pereira, Henrike Häbel. In addition to the great people at the Department of Mathematics at Chalmers, I really must thank all the amazingly nice people at Sahlgrenska Academy that a couple of years ago got me thinking that an academic career might be fun. Thank you: Anna Johnning, Johan Bengtsson-Palme, Filip Cuklev, Joakim Larsson, and Lina Gunnarsson Kearney.

Life and work would not be great without all the things that happen outside of the workplace either, all of which contribute to a very happy existence. My fellow band members of DÖDAREN: Jon Solheim, Martin Claesson, Erik Rådegård, without you life just would not be as fun. Josefin Westin and Ida Backström should consider themselves included in the previous statement!

I am happy to have great family members that always support me in whatever I do. Thank you Jan, Rie, Clara, Vera, and Ulrika!

Last but in no way the least, Magdalena Isfelt, your name deserves to be in here and you know why.

List of Papers

The licentiate thesis includes the following papers.

I. **F. Boulund**, A. Johnning, M.B. Pereira, D.G.J. Larsson, E. Kristiansson (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. C-F. Flach, **F. Boulund**, E. Kristiansson, D.G.J. Larsson (2013). Functional verification of computationally predicted qnr genes. *Submitted.*

III. **F. Boulund**, A. Sjögren, E. Kristiansson (2013). A framework for distributed gene quantification. *Manuscript*

Below follows a list of relevant papers that are not included in this licentiate thesis.

- J. Bengtsson-Palme, **F. Boulund**, B. Weijdegård, C-F. Flach, J. Fick, E. Kristiansson, D.G.J. Larsson (2013). Metagenomics reveal a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Manuscript.*

- C.S. Hakimi, C. Hesse, H. Wallén, **F. Boulund**, A. Grahn, A. Jepsson (2013). Assessment of platelet storage lesion with multiple electrode aggregometry. *Submitted.*

Contents

Chapter 1

Introduction

Bioinformatics is a multidisciplinary field of research at the intersection of biology, computer science, mathematics, and statistics. Because of this combination of disciplines there might be a lot of ground to cover before a person unversed in all of these can appreciate the issues at hand. This chapter aims to describe a minimal background so that the reader is equipped with some of the prerequisite knowledge to understand the basic underlying concepts in the research papers included in the thesis. The articles concern topics ranging from the identification of novel antibiotic resistance genes in the environment using hidden Markov models, to utilizing high-performance computing clusters to quantify gene content in order to compare microbial communities. To introduce the reader to this relatively wide background the introduction will cover the essentials of DNA, DNA sequencing, metagenomics, and antibiotic resistance. Furthermore, it will try to highlight why large computer clusters are sometimes required to perform analyses of metagenomic DNA sequence data.

This thesis contains three research articles. The first two articles have a natural relation to each other; the second makes biological verifications of some of the results that were discovered in the first. This relation highlights a modern way of doing research where hypotheses are generated using computational methods (i.e. bioinformatics) and then verified in the lab, rather than other way around which has historically been more common. The third article included in the thesis describes a method that leverages the parallelism of high-performance computer clusters to perform analyses of very large metagenomic data sets being produced today and in the future.

1.1 DNA

Deoxyribonucleic acid (DNA) is a biological molecule. It is used in all living organisms for storing genetic instructions (cf. a blueprint). It is a stable molecule that is copied in cell division processes to transfer genetic information (i.e. inheritance). The molecule consists of basic building blocks called nucleotide bases which are connected together to form almost arbitrarily long strands of DNA via a sugar backbone (deoxyribose). There are four kinds of nucleotides, normally abbreviated to single letters as such: adenine (A), cytosine (C), guanine (G), thymine (T). The nucleotide bases also have the ability to form pairwise bonds; A with T, and C with G. This is called *base pairing* and is what enables DNA to form helical structures when two long complementary strands of DNA base pair to each other (Watson and Crick, 1953). This is in part why DNA is such a stable molecule and well suited for carrying the important genetic information of an organism.

An important concept in molecular biology is *The Central Dogma* (Crick, 1970), which states how information flows in biological systems. The key point is that biological information is encoded in the DNA as sequences of adjacent nucleotides. Throughout this thesis, I will refer to such regions as *genes*. Within a gene, each sequence of three consecutive nucleotides encode for an amino acid. There are 20 different standard amino acids that are encoded for in normal DNA. The genes are translated through a number of processes into long chains of amino acids (cf. a pearl necklace). Such chains of amino acids fold together to become proteins. Proteins are large biological molecules and one of the major constituents of every cell. They perform a wide variety of functions in the cells, some of which will be further detailed in section 1.4. The central dogma is vital for realising that proteins can be identified in DNA sequences and (to some extent) vice versa. This relation is essential for all research papers included in the thesis. For example, one of the concepts underlying Paper I 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.

1.2 DNA sequencing

In order to investigate the sequence of the DNA of an organism, the actual DNA molecule(s) in that organism needs to be sequenced. That means that the DNA inside the organism must be extracted and the order of the nucleotide bases determined. In 1975 a method now called Sanger sequencing (Sanger, 1977; Sanger and Coulson, 1975) was published. Modern variants of this method are in some cases still used because of its merits in sequencing relatively long continuous stretches of DNA (700-900 bases) at a high quality. The throughput of Sanger sequencing is however far too low to be time and cost effective in a majority of scenarios, making it used mostly for smaller sequencing projects.

About a decade ago there was a revolution in DNA sequencing as the *next-generation sequencing* (NGS) era began when novel methods for high-throughput DNA sequencing started becoming commercially available (Stein, 2008). It brought with it a wide use of a now common method called whole genome shotgun sequencing. In this method DNA is extracted from samples of interest and enzymatically broken into many shorter pieces. These short pieces of DNA are put into a sequencing machine that determines what nucleotides are present and in what order. One of the hallmarks of NGS technologies is that all fragments are sequenced in parallel, leading to the early designation "massively parallel sequencing".

The output from NGS machines is called *reads*. Each read represents a short section of DNA; historically machines produced reads as short as 25-50 bases, but today's current technologies can produce read lengths between 100-250 (some up to 400+ bases), depending on what technology is used. Several different NGS technologies exist which can be grouped into several categories. The most common technologies that are currently available in the market and 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 more recently single-molecule sequencing (e.g. Pacific Biosciences SMRT).

A majority of all the sequencing data used in this thesis was produced using Illumina sequencers. They implement the sequence by synthesis approach to DNA sequencing which starts with sample preparation where the DNA is extracted, purified, and fragmented into *templates*. These 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 flow cell of an Illumina HiSeq 2500 can have between 610,000-678,000 such clusters per square millimeter (Illumina, 2013), which is one of the reasons for their immense throughput. The actual sequence of nucleotides in the strand clusters is determined by flowing a solution of free nucleotides across the flow cell. The nucleotides attach to their complementary base on the strands, and 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 light emitted, coded with different colours for each of the four nucleotides, to produce a sequence of images. From these, the sequence of nucleotides on the strands in each cluster can be determined. This constitutes the *base calling* process. For interested readers a very good presentation of different NGS technologies is presented in (Metzker, 2010).

Because the DNA sequence data produced by NGS machines consist of short reads it requires special techniques for analysis. Often the first step in an analysis is to *assemble* the reads, like a massively big and complex jigsaw puzzle, before any further analysis is made. Unfortunately, the sequencing machines are not perfect and occasionally return erroneous base calls (Treangen and Salzberg, 2012). Different technologies are more prone to different types of errors. In 454 pyrosequencing, insertions occasionally mistakenly occur (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 (deletions). Illumina on the other hand, is more prone to miscall bases, leading to incorrect nucleotides 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). Regardless of the type, errors introduce complexities in the analysis (Meacham et al., 2011). They can make it harder to correctly place from where the read originates in a reference, or severely complicate assembly, making it important to consider the occurrence of errors in sequence data. There has been a lot of research effort put into trying to model the occurrence of errors in reads and trying to correct them, as highlighted by the vast numbers of available error correction algorithms, some examples of which include the commonly used FASTX (Gordon and Hannon, 2010), and HTQC (Yang et al., 2013), as well as 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).

One of the main advantages of modern NGS technologies is that they produce a lot more sequence information in total than earlier methods. This turns out to be both a blessing and a curse. Current NGS platforms such as Illumina can produce as much as several gigabases per run (10^9 nucleotide bases). Recent and coming data sets produced by consortia such as Meta-HIT (Qin et al., 2010), The Earth Microbiome Project (Gilbert et al., 2011), and The Human Microbiome Project (Turnbaugh et al., 2007) are expected to deliver data sets on the order of petabases (10^{15} bases) over the next few years (Baker, 2010). DNA sequence data storage organisations like the European Nucleotide Archive describe that they experience a doubling of sequence data every 10 months (Cochrane et al., 2013), a rate that is higher than the expected growth in e.g. harddrive sizes as predicted by Kryder's Law (Walter, 2005) and also surpasses the predictions of Moore's law (Schaller, 1997) that relates to computer processing power. This means that to analyse data sets of such massive sizes will require novel methods and approaches, because the currently employed algorithms and computational methods will no longer be capable of handling the coming data deluge.

1.3 Metagenomics

Metagenomics is a subject central to this thesis. It concerns the examination of genetic material extracted from environmental samples. Much like a genome is the complete genetic material of a single organism, a metagenome is the collective genetic material of an entire community of organisms (e.g. bacteria). The term was coined in 1998 (Handelsman et al., 1998) to describe the application of genomic techniques to uncultured microorganisms. It has been shown that less than 1% of environmental bacteria are possible to grow in the lab (Hugenholtz et al., 1998). This would leave the remaining 99% of environmental bacteria out of reach if culture independent methods like metagenomics did not exist. Next-generation sequencing is the enabling technology that allows for the study of the genetic makeup of such microbial communities.

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 de facto standard in large-scale metagenomic sequencing studies. Unfortunately, as a result of how modern whole genome shotgun NGS technology works, the sequence data from a metagenomic sample consists of a massive amount of of short reads with all the genetic material in the sample mixed together. This makes it difficult to determine (or piece together) exactly what organisms were present in the sample, and in what quantities. In addition, the DNA of highly abundant species in an environmental sample are overrepresented in the sequenced DNA because of the random selection of DNA in shotgun sequencing, due to their relative abundance. To determine the features of the genomes of less abundant organisms a high coverage is required. Thus large amounts of DNA needs to be sequenced to ensure that the less abundant organisms have DNA fragments represented in the metagenomic sample. However, because of the immense output of NGS machines the size of a large modern metagenomic study can easily reach several gigabases, many even approaching terabases. In fact, the size of the data sets is in itself creating the need for more focused research efforts on making data analysis more efficient.

Nowadays metagenomics is often associated with, and thought of, as a way to study bacterial communities. Among the first uses of metagenomics was however to study the presence of viral DNA in seawater. Seawater was also the focus of Craig Venter's ambitious Gene Ocean Sampling (Yooseph et al., 2007; Williamson et al., 2008) project in which a sailing boat was equipped with a DNA sequencing lab and set off to sail across the globe to sample seawater for DNA. It was a pioneering project in the investigation of environmental metagenomes and helped mature the field, bringing many improvements in data management and large-scale data analysis with it.

An application where metagenomics is often used is when trying to deduce differences between samples or communities (Jones et al., 2010; Tringe et al., 2005). In order to perform such comparisons it is necessary to quantify the presence of genes or other features of interest (Delmont et al., 2013). Quantification enables comparison between different samples, for example to compare the content of resistance genes in the intestinal microbiota of healthy individuals not exposed to antibiotics with that of people with chronic intestinal diseases that are dependent on regular use of antibiotics. To perform gene quantification requires that each read in the metagenomic sample is compared to a reference (e.g. databases of known antibiotic resistance genes in 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 comparison (Kristiansson et al., 2009). Sequence comparison (i.e. sequence alignment) is a task of relatively high computational complexity, both requiring clever algorithms and powerful computer hardware to perform efficiently at the scale required for large modern metagenomic data sets.

1.4 Antibiotic resistance

Paper I and II uses the previously described techniques to investigate the presence of antibiotic resistance genes in environmental bacterial communities. Antibiotics are substances, chemical or biological, that either prevent growth of, or even kill, bacteria. Modern health care is therefore highly reliant on effective antibiotics to treat bacterial infections (Rosenblatt-Farrell, 2009). However, bacteria are becoming resistant to many of our commonly used antibiotics (Neu, 1992) and many argue for careful usage of antibiotics (Andersson and Hughes, 2010, 2012).

Microbial organisms have a long history of withstanding antibiotics (D'Costa et al., 2011; Davies and Davies, 2010; Sykes, 2010). At their disposal is a large toolbox of methods available for their use Allen et al. (2010). Some examples of how bacteria can acquire antibiotic resistance include: *a*) modification/alteration of the active substance, e.g. β-lactamases that break down the β-lactam antibiotics (penicillin is a common β-lactam antibiotic), *b*) alteration of the target site, e.g. fluoroquinolone resistance by mutation of the amino acid sequence of the protein gyrase that the antibiotic targets, *c*) transport of the active substance out of the cell (e.g. using efflux pumps). Often these mechanisms are available in the genome of the organism, but bacteria also have several mechanisms with which they can exchange genetic material with other bacteria through a process called horizontal gene transfer (Aminov and Mackie, 2007; Bennett, 2008). This enables bacteria to share genes for antibiotic resistance. The transfer often occurs via ring shaped DNA constructs called plasmids (Salyers et al., 2004). Each plasmid can contain several genes and sometimes even the instructions (genes) for how to construct the machinery required for the transfer (Rajpara et al., 2009; Jacobsen et al., 2007; Hall and Collis, 1995).

Many common bacterial antibiotic resistance genes originate from the environment (Allen et al., 2010; Cantón, 2009). There are several examples of metagenomic sequencing projects that has focused on different types of microbial communities, e.g. The Earth Microbiome Project (Gilbert et al., 2011, 2010), and The Human Microbiome Project (Blaser, 2010; Turnbaugh et al., 2007), as well as several others (Penders et al., 2013; Qin et al., 2010; Huttenhower et al., 2012; Lazarevic et al., 2009; Jones et al., 2010; Eckburg et al., 2005; Gill et al., 2006). With metagenomic and computational techniques it is possible to identify antibiotic resistance genes in environmental microbiomes (Schmieder and Edwards, 2012). The availability of such metagenomic data sets enable the study of known antibiotic resistance genes in such environments, but also provides the possibility of finding new, previously undiscovered antibiotic resistance genes.

Chapter 2

Introduction to papers

This chapter consists of brief introductions to the three papers included in this thesis. The introductions provide an alternative, more accessible, angle to the papers than their respective abstracts in order to provide unversed readers with a basic introduction that can be further expanded by reading the abstracts and the papers. I also try to highlight the main findings or main conclusions from each paper to further make it easy for the reader to understand their respective contributions.

2.1 Introduction to Paper I

In the first paper, *A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences*, we developed a technique that enables the identification of a special type of antibiotic resistance genes called *qnr*, in short fragments of DNA. The *qnr* genes are a family of plasmid mediated resistance genes that provide bacteria with resistance to commonly used antibiotics in the fluoroquinolone family (e.g. ciprofloxacin). The method we developed is based around the fact that *qnr* proteins have a specific repeat pattern in their amino acid sequence, lending them the classification *pentapeptide repeat proteins.* As such, their sequence can be accurately discriminated using a statistical modelling framework called hidden Markov models. Such models are able to capture the repeating pattern in the amino acid sequence while still allowing for high variability in the regions where there is little conservation of amino acids between the different gene family members.

The main findings of Paper I are that the *qnr* family of antibiotic resistance genes are very suitable to describe using a hidden Markov model. Such a model can is able to accurately identify the amino acid sequence of a *qnr* fragment and distinguish it from other similar pentapeptide repeat sequences that lack a resistance phenotype. In our study we also discovered several fragments that indicate a presence of previously unknown *qnr* gene variants in the environment. This paper was published in *BMC Genomics 2012* **13**:695. (Boulund et al., 2012).

2.2 Introduction to Paper II

The second paper, *Functional verification of computationally predicted qnr genes*, is a follow up study of the results from Paper I. The paper describes the development of a bacterial expression platform for evaluation of the resistance phenotype of antibiotic resistance genes. The chosen host organism was *Escherichia coli* and the expression platform was evaluated using synthesized genes of several types of well-known *qnr* genes as well as some of the novel candidates we discovered in Paper I.

Paper II shows that two putative *qnr* genes discovered in Paper I do provide antibiotic resistance when expressed in an *E. coli* host. The experimental work was performed by collaborator Carl-Fredrik Flach from The Sahlgrenska Academy at University of Gothenburg. It is also an example of how computational methods can be used in exploratory studies to generate novel hypotheses and intermediary results that can be verified in the lab, rather than the other way around which has traditionally been more common. This paper has been submitted to a peer-reviewed scientific journal.

2.3 Introduction to Paper III

In the third paper, *A framework for distributed gene quantification*, we addressed the challenge of working with very large scale metagenomic data sets. Since the size of metagenomic data sets are continually increasing and this increase is showing no signs of slowing down, novel methods for analysing these quantities of data are required. 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).

The fundamental result of Paper III is a new method for distributed gene quantification in very large metagenomic data sets. The method is based on a bioinformatics pipeline consisting of the components required to perform gene quantification in metagenomic data, which is then distributed across several nodes in high-performance computer clusters in a data parallel manner, exploiting the independence and inherent possibilities for parallelism in read mapping. This is a paper in manuscript form, expected to be submitted for peer-review later this year.

Chapter 3

Future work

A natural continuation of the work in Paper I is to apply the method to more recent and larger data sets. This work has already begun with a project student that I co-supervised during the summer of 2013. We will focus on continuing to improve the throughput of the pipeline to allow for terabase-scale short read data. In addition, the method has been evaluated for use in scenarios other than the identification of *qnr* genes, e.g. with β-lactamases, to further broaden the utility of the method.

There are several improvements to the framework described in Paper III currently in progress. Some details to improve the overall ease-of-use is still required. This includes the production of a solid API reference and tutorial material to enable widespread usage of the framework. One major missing feature of the framework is support for blastx-like functionality to enable mapping to protein databases using unmodified FASTQ read data. As the framework reaches maturity and further optimizations to parts of the pipeline has been made the ambition is to make it capable of handling input data sizes up to 10 terabases.

Next-generation sequencing has enabled whole community analysis of the genetic content of environments such as the human gut. As mentioned in section 1.4, there is a great interest in studying the complex bacterial communities that occupy this habitat. A pilot project studying the antibiotic resistance gene content in the human gut was performed by me in my master thesis project (Boulund, 2010). With the availability of the pipeline described in Paper III it is now possible to further investigate the presence of antibiotic resistance genes in the large gut metagenomes published in recent years. This work will continue in collaboration with people from The Systems and Synthetic Biology group at the Department of Chemical and Biological Engineering at Chalmers University of Technology.

In addition, our ambition is that the method from Paper III will find itself applied in several collaborative efforts where quantification of metagenomic data is required. One such project is NICE (Novel Instruments for effect-based assessment of Chemical pollution in coastal Ecosystems; `www.nice.gu.se`) which studies how marine microorganisms are affected by pollution. The effects will be studied, in part, by comparing marine samples of polluted and pristine environments using comparative metagenomic techniques. Another project where the method is well suited and expected to be applied is in collaboration with members of the INTERACT project (`www.interact.gu.se`), where bacterial resistance to metals and antibiotics is studied.

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.

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.

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

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, 153 Suppl 1(January):S347–57.

Blaser, M. J. (2010). Harnessing the power of the human microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6125–6.

Boulund, F. (2010). *Exploring antibiotic resistance genes in the human intestinal microbiome.* Master of Science Thesis, Chalmers University of Technology.

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

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.

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):D30–5.

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., 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):457–61.

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.

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.

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.

Gilbert, J. a., Bailey, M., Field, D., Fierer, N., Fuhrman, J. a., Hu, B., Jansson, J., Knight, R., Kowalchuk, G. a., Kyrpides, N. C., Meyer, F., and Stevens, R. (2011). The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2011. *Standards in Genomic Sciences*, 5(2):243–247.

Gilbert, J. a., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. T., Brown, C. T., Desai, N., Eisen, J. a., Evers, D., Field, D., Feng, W., Huson, D., Jansson, J., Knight, R., Knight, J., Kolker, E., Konstantindis, K., Kostka, J., Kyrpides, N., Mackelprang, R., McHardy, A., Quince, C., Raes, J., Sczyrba, A., Shade, A., and Stevens, R. (2010). Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Standards in genomic sciences*, 3(3):243–8.

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.

Gordon, A. and Hannon, G. (2010). Fastx-toolkit, fastq/a short-reads preprocessing tools (unpublished). `http://hannonlab.cshl.edu/fastx_toolkit`. [Online; accessed 2013-September-09].

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.

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):R245–9.

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.

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.

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.

Illumina (2013). Performance and Specifications for HiSeq 2500/1500. `http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn`. [Online; accessed 2013-September-09].

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.

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.

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

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osterå s, 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.

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.

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.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46.

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

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.

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. r., 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. r., 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.

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.

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.

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

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

Stein, R. (2008). Next-generation sequencing update. *Genetic Engineering and Biotechnology News*, 28(15).

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

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.

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.

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

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.

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.

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.

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.