Bayesian mixture models for metagenomic community profiling

Sofia Morfopoulou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

UCL Genetics Institute

Department of Genetics, Evolution and Environment

University College London

December 21, 2015

I, Sofia Morfopoulou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Metagenomics can be defined as the study of DNA sequences from environmental or community samples. This is a rapidly progressing field and application ideas that seemed outlandish a few years ago are now routine and familiar. Metagenomics' scope is broad and includes the analysis of a diverse set of samples such as environmental or clinical samples. Human tissues are in essence metagenomic samples due to the presence of microorganisms, such as bacteria, viruses and fungi in both healthy and diseased individuals.

Deep sequencing of clinical samples is now an established tool for pathogen detection, with direct medical applications. The large amount of data generated produces an opportunity to detect species even at very low levels, provided that computational tools can effectively profile the relevant metagenomic communities. Data interpretation is complicated by the fact that short sequencing reads can match multiple organisms and by the lack of completeness of existing databases, particularly for viruses.

The research presented in this thesis focuses on using Bayesian Mixture Model techniques to produce taxonomic profiles for metagenomic data. A novel Bayesian mixture model framework for resolving complex metagenomic mixtures is introduced, called metaMix. The use of parallel Monte Carlo Markov chains (MCMC) for the exploration of the species space enables the identification of the set of species most likely to contribute to the mixture. The improved accuracy of metaMix compared to relevant methods is demonstrated, particularly for profiling complex communities consisting of several related species. metaMix was designed specifically for the analysis of deep transcriptome sequencing datasets, with a focus on viral pathogen detection. However, the principles are generally applicable to all types of metagenomic mixtures. metaMix is implemented as a user friendly R package, freely available on CRAN: http://cran.r-project.org/web/packages/metaMix.

Acknowledgements

I would like to first thank my supervisor Vincent Plagnol for giving me the opportunity to carry out my postgraduate studies with him in UCL, in a stimulating research environment. Thank you for your support and guidance over the years.

I would also like to thank my second supervisor David Balding for his thoughtful advice and feedback during crucial moments of the PhD.

A more recent professional connection I want to mention is with Judy Breuer, I feel lucky for the opportunity to see my work applied in a real clinical setting.

I am very glad to have met all my friends in UGI: you made sure there was never a dull moment in the office (OK there were some!). A special thank you goes to Claudia, my closest ally at work. I am happy we went through the PhD in parallel, you always made me laugh during frustrating moments (A++ for the timely use of some greek phrases). I enjoyed a lot the interaction with the rest of the UGI gang: Vale, Doug, Delilah, Julie and also Jon, Warren and Cian and of course everyone else - thank you all.

My non-UGI friends in London played a big part towards maintaining a relatively balanced life during these past four years. Special mentions have to go to Tasos, Giorgos and Karolina. I owe a lot of my well-being to my friends back home - well back home is not accurate as most of you are scattered around the world - but Peni, Anna, Artzi, Andrea and Vaggeli, they say that in university you find ”your people” and this was definitely the case for me. Conversation never flows more easily than when we are together. Silliness and summer camping for life!

To my parents Maria and Dimitris and my sisters Alexandra and Myrto: I love you a lot, you have always respected my choices and helped me towards making these true. I dedicate the thesis to you. I hope it will not be too long before we live closer to each other. I also want to mention Tasos, you are a great dude plus you are half responsible

for Marilena, who is the best person on earth. Marilenaki you are my favorite of them all and I miss you a lot. Also a big hug to Elias, Lila and of course Jason and Natalia, my very handsome extended family.

And finally Michalis: your love gives me courage and makes me feel free. I have the most fun with you, here's to our new adventures!

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

During this PhD project, the problem that interested me the most was how to efficiently perform sensitive community profiling in a metagenomics sample. Differently stated, this is the fine-grained identification of species present in a metagenomics sample, coupled with the ability to find the "needle in a haystack".

Community profiling is an active research question and a substantial amount of work has been produced towards answering it (Huson *et al.*, 2007; Xia *et al.*, 2011; Segata *et al.*, 2012; Francis *et al.*, 2013; Wood and Salzberg, 2014). While these methods have been widely used in practice, there are some yet unaddressed limitations. For example, many methods perform taxonomic assignment for each read individually, ignoring the information provided by the rest of the data. Furthermore, more complex models typically fit the data better and methods that ignore this known problem are destined to infer increasingly complex profiles, exhibiting low specificity by introducing a significant number of false positives species. Mixture models (McLachlan and Peel, 2000) can help with the first issue, while Bayesian methods (Jeffreys, 1961; Gull, 1988) address the second. In this PhD project I worked towards developing a new framework for community profiling by taking a Bayesian mixture model approach. This also provides a coherent way to estimate the probability of a species being present as well as the read assignment probability.

To illustrate the context of the thesis I start with setting the terminology for metagenomic research. I then provide a historical perspective in molecular biology and the relevant discoveries and advances in technology that make metagnomics research feasible today. An introductory overview of diagnostics and metagenomic diagnostics follows and I discuss the ways high throughput sequencing has reshaped it. Subsequently, the community profiling problem is defined and prior relevant methodological work is discussed, including the most popular approaches. The relevant challenges and limitations are highlighted. The bioinformatics analysis required as a prior step to species identification and quantification when dealing with deep sequencing data generated from human clinical samples is outlined. Finally, the requirement for a Bayesian mixture model-based solution is emphasized and the ideas underlying the novel proposed method called metaMix are introduced.

1.1 Microbiome definitions and a few words on viruses

For most of earth's history, life consisted solely of microscopic life forms and microbial life still dominates the planet in many aspects. The collection of microorganisms that occupy various sites of the human body is called the human microbiota (Marchesi and Ravel, 2015) and it includes viruses, bacteria and fungi. The definition of the term microbiome has a convoluted history. My personal preference is to use the word microbiome to refer to the set of resident microorganisms and associated abiotic factors of given environments (Marchesi and Ravel, 2015). However the same term has been used to either describe just the population of microbes that colonise the human body (Petrosino *et al.*, 2009) or even further limited to define the complete set of genetic information associated with a set of microorganisms (Matsen, 2015). The latter definition in my opinion best describes the metagenome. The microbiome can be characterized by different approaches such as metagenomics (Blomström *et al.*, 2010; Ng *et al.*, 2011), metatranscriptomics (Santos *et al.*, 2011) or their combinations. Metagenomics is thus the genomic study of uncultured microorganisms living in environmental niches, plants or animal hosts (Chen and Pachter, 2005) while metatranscriptomics is the analysis of RNA sequence data from such samples.

The launching of international human microbiome projects (such as `http://hmpdacc.org/`) highlights the significance of understanding the microbiome. The metagenomics field has been progressing rapidly and application ideas that seemed outlandish a few years ago are now routine and familiar. There are many exciting applications that require the analysis of a diverse set of samples such as gut microbiome (Qin *et al.*, 2010; Minot *et al.*, 2011), environmental (Mizuno *et al.*, 2013) or clinical (Willner *et al.*, 2009; Negredo *et al.*, 2011; McMullan *et al.*, 2012) samples.

Among these applications, the discovery of viral pathogens is relevant for clinical practice (Fancello *et al.*, 2012; Chiu, 2013). Viruses are recognised to be the most abundant biological entities on the planet (Breitbart and Rohwer, 2005). Descriptions of viral infections appear throughout human recorded history and long before viruses were first discovered. Viruses can cause diseases in plants, animals and humans; however, healthy individuals are also chronically infected by a number of viruses without any detectable symptoms (Virgin *et al.*, 2009). Viruses and bacteria are known to play a role in the pathogenesis of various human diseases. Studying the human metagenome is thus highly relevant to understanding infectious as well as common complex diseases. The focus of the work undertaken for this thesis and my main interest was to develop a method for community profiling and read interpretation from metagenomic data, with a particular focus on viral detection in human clinical samples. The introduction is therefore built around a viral core and the following sections will cover in greater depth the virus-related aspects of this work.

1.2 Molecular biology: a historical perspective

Prior to an exploration of the metagenomics field, we provide an overview of the history of molecular biology to better appreciate the strengths and limitations of the technological advances that revolutionised it.

Molecular biology encompasses all research on the structure, function and interactions of biological macromolecules. This includes research on the molecular nature of the gene and the mechanisms of gene replication, mutation, and expression (Morange, 2009). Most historians of biology agree that the origins of molecular biology can be traced before the World War II (Morange, 2009) when different technologies such as electrophoresis, X-ray crystallography, electron microscopy were introduced. These resulted in the initial discoveries on the macromolecules structure and paved the way for seminal research accomplishments.

1.2.1 Classical era of molecular biology

The classical era of molecular biology roughly spans the period between 1940 and 1965, during which the field blossomed. A first discovery of major scientific importance identified the deoxyribonucleic acid (DNA) as the major constituent of the genes

(Hershey and Chase, 1952). The experiment used phage viruses to confirm that the genetic material transmitted from generation to generation is DNA and not proteins.

One of the most famous advances, still celebrated today, is the discovery of the double helix structure of DNA composed of four bases (Figure 1.1) by James Watson and Francis Crick (Watson and Crick, 1953a,b). Watson and Crick presented a model of the double helical structure of DNA, where the complementary strands are held together by hydrogen-bonded base pairs. The two collaborators used extensively the X-ray crystallography work on DNA by Maurice Wilkins and Rosalind Franklin. In their second article in Nature a month later, Watson and Crick presented DNA as a genetic information molecule.

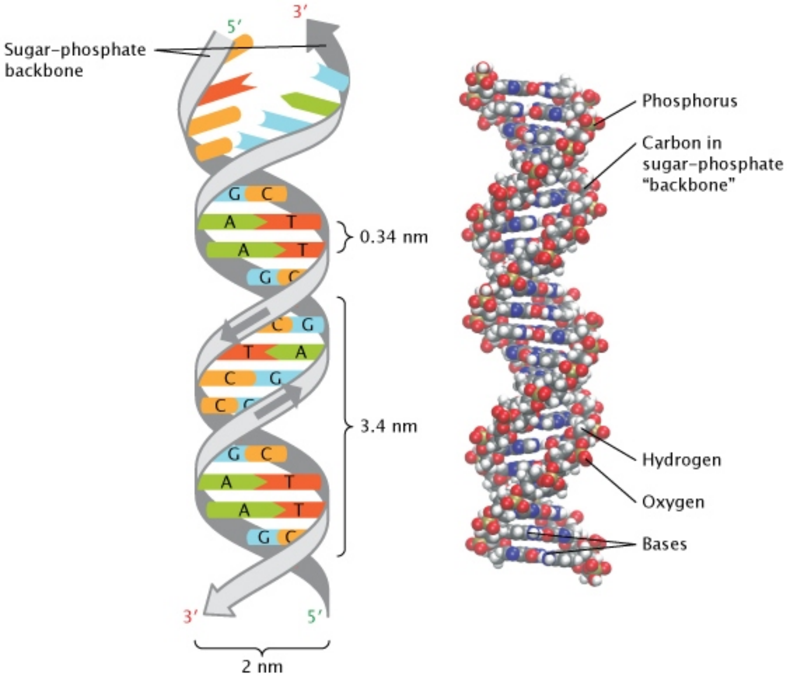


Figure 1.1: The DNA double helix structure: complementary bases are held together as a pair by hydrogen bonds. The figure is taken from (Pray, 2008).

The double helical structure discovery influenced and shifted contemporary research in molecular biology towards understanding the genetic replication and function mechanisms. The now famous central dogma of molecular biology according to which information flows from the nucleic acids to the proteins and only in this direction was proposed a few yeas later by Crick:

> “This states that once information has passed into protein it cannot get

> out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein." (Crick, 1958)

Subsequent discoveries challenged the concept of a linear relationship between a DNA sequence and the produced protein. Findings included the obervation that DNA consists of coding regions (exons) interspersed with non-coding regions (introns) and that the exons may be separated by long stretches of non-coding DNA. Additionally, different exons may be spliced together thus generating a variety of molecular products. Gradually, molecular biologists started to update their understanding of what constitutes a gene.

1.2.2 Genomic era and landmarks in DNA Sequencing

The first succesful attempt at sequencing took place in the mid 1960s with the characterisation of the complete sequence and structure of an RNA molecule (Holley *et al.*, 1965) Potentially the most significant contributions towards sequencing DNA molecules were by Frederick Sanger who established elegant DNA sequencing techniques in the 1970s. Sanger sequencing, an enzymatic method using DNA polymerase, was published in 1975 (Sanger and Coulson, 1975). An easier and more efficient chain termination method that employes fluorescently labeled dideoxynucleotides (ddNTP) for chain termination was published two years later (Sanger *et al.*, 1977a). The first ever DNA genome to be fully sequenced was ΦX174 (Sanger *et al.*, 1977b), a bacteriophage that infects *Escherichia coli*. Subsequently, the chain terminator method was rapidly employed and it was the technological platform conventionally used in genomic and metagenomic studies up until the arrival of high-throughput sequencing.

Two significant developments further advanced the sequencing field in the 1980s. The first one was the polymerase chain reaction technique or PCR, a DNA amplification technique (Mullis and Faloona, 1987; Saiki *et al.*, 1988). PCR is the method where a nucleic acid sequence is exponentially amplified in vitro through a polymerase-catalyzed chain reaction. The second was the development of automated DNA sequencing instruments by Applied Biosystems (reviewed in (Liu *et al.*, 2012)).

The International Project on Human Genome was initiated in 1990 and it was expected to last 15 years. In 2000 after ten years of multinational scientific effort and at $3 billion cost, a rough draft of the genome was finished using the Sanger sequencing method with key findings of the draft genome announced soon afterwards (Lander *et al.*, 2001).

1.3 Deep sequencing technology

Deep sequencing has been a groundbreaking technology, affecting the whole breadth of the biomedical sciences. The key feature is the potential for massive parallelisation and automation, making large scale sequencing projects possible. It has had a huge impact first on genomics and soon afterwards on metagenomics as it provides the opportunity to sequence uncultured microorganisms sampled directly from their natural habitats.

Since its arrival, intense competition between the major players has contributed to the constant improvement of the technology. This is specifically expressed by the continuous increase in numbers and lengths of reads, translating into a reduced cost per sequenced base. The most popular high throughput platforms are Illumina (Bentley *et al.*, 2008), Roche 454 (Margulies *et al.*, 2005), IonTorrent (Rothberg *et al.*, 2011), SOLiD (Shendure *et al.*, 2005) and Pacific Biosciences (Eid *et al.*, 2009). Despite the continuous improvement in performance over the last years for all platforms, there is an important variation of throughput and read length between them. The output of Illumina, SOLiD and Ion Torrent consists of reads at most a few hundred bases long, while Pacific Biosciences generates longer reads, kilo-bases long.

During the first years of metagenomic studies, the Roche 454 was the favored platform for sequencing metagenomes due to the longer read length. With longer reads taxonomic assignment of the reads is relatively easy when the reference genomes are known. However due to the relatively limited throughput, rare species were habitually missed. On the other hand, Illumina platform generates an order of magnitude more reads of reduced length at a lower cost, which increases the chance of identifying low abundance species. Illumina currently dominates the market with various machines and offer the best balance between read lengths, error rates and cost (Loman *et al.*, 2012). The trend of rapid turn-around time and falling cost supports a prediction that once high throughput sequencing become widely accessible, its use as a clinical diagnostic tool will allow more personalized medical applications. Especially relevant to medical applications is the technology developed by Oxford Nanopore Technologies (ONT). ONT has given early access of their mobile USB-powered single molecule sequencer to a number of academic collaborators for evaluation. Even though a commercial launch has not been announced yet, assuming that ONT deliver on the low-cost portable sequencer producing very long sequences with low error rates on the spot, there is great potential for its use in the emerging clinical market.

1.3.1 Illumina Sequencing

Illumina sequencing will be discussed in more detail, as it has been used for the entirety of the sequencing carried out for this thesis. Illumina sequencing was first available in 2006. It is based on a sequencing by synthesis approach where a polymerase is used to synthesise a complementary strand to the single stranded target DNA with terminator nucleotides used to halt the synthesis. The terminators are reversible which means that the synthesis can continue after each base is detected. The sequencing process is performed for millions of DNA fragments in parallel. In brief, there are three main components in the process: library preparation (Figure 1.2), bridge amplification (Figure 1.3) and sequencing by synthesis (Figure 1.4).

During the library preparation step (Figure 1.2), the DNA is fragmented and tagmented. The aim is to generate overlapping fragments within a specific size range. Short nucleotide sequences called adaptors are attached to the template fragments and they serve multiple purposes: they include complementary oligos that allow the fragment to ligate to the flowcell[1], index tags to label the sample and allow multiplexing and binding sites for the sequencing primers. Typically after adapter ligation, a PCR step is used to enrich the library for DNA fragments with the adaptors in the correct orientation. The DNA is then denatured to produce single strands.

The single-stranded sequencing library is loaded into the flow cell, followed by bridge amplification (Figure 1.3). This is an amplification reaction that occurs on the surface of the Illumina flow cell. The library fragments bind to complementary oligos as they “flow” across the oligo lawn while the opposite end of a ligated fragment bends

[1] flowcell: a glass slide with one, two, or eight physically separated lanes, depending on instrument platform. Each lane is coated with a lawn of surface bound, adapter-complimentary oligos.

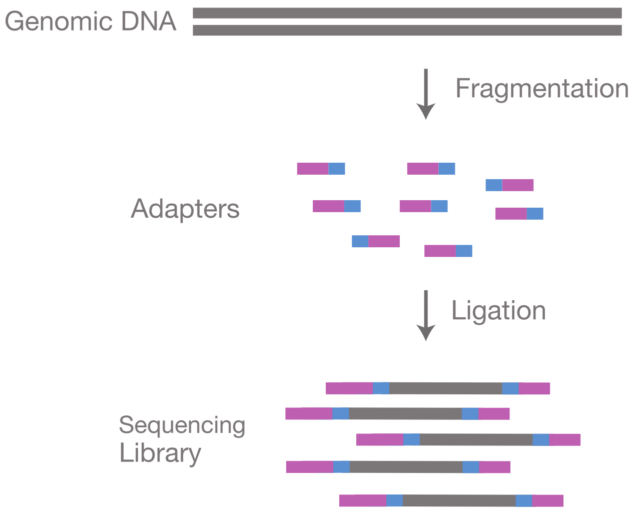


Figure 1.2: Library preparation: the library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends. The figure is taken from the Illumina website.

over to the surface, bridging on a complementary oligo. The repeated denaturation and extension cycles for each single fragment results into millions of dense clusters of clonal DNA across the flow cell. This in turn, increases the fluorescent signal intensity.

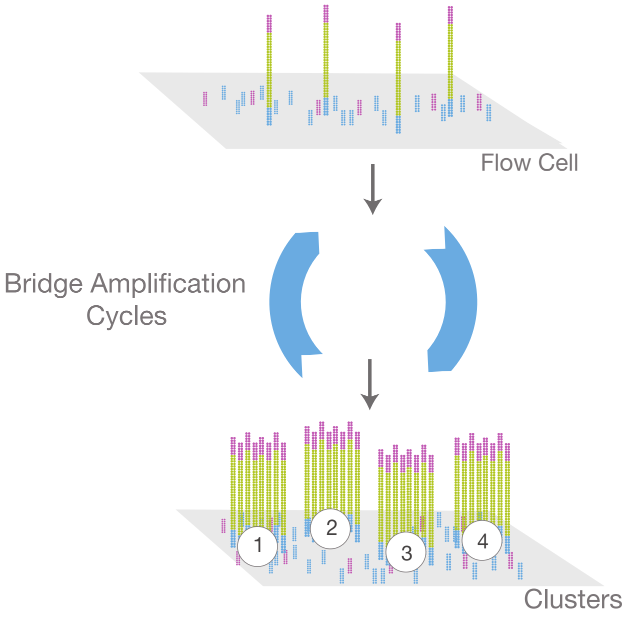
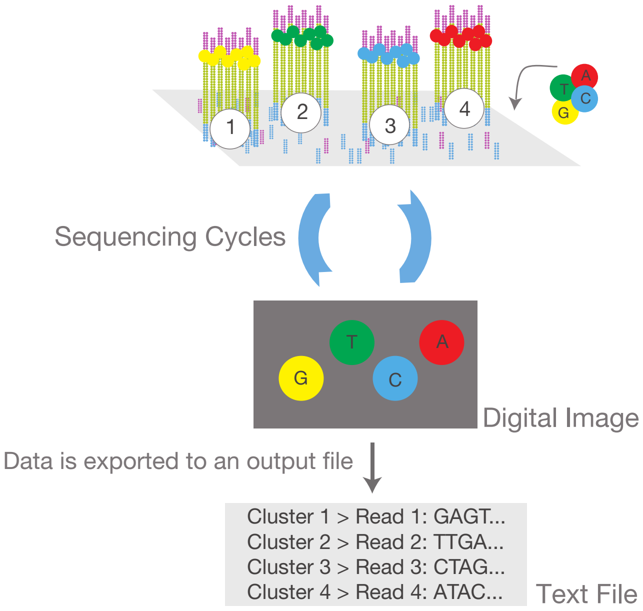


Figure 1.3: Bridge amplification: the library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification. The figure is taken from the Illumina website.



Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...

Figure 1.4: SBS sequencing: Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated n times to create a read length of n bases. The figure is taken from the Illumina website.

The sequencing reaction (Figure 1.4) is carried out with fluorescently labeled reversible terminator-bound dNTPs (modified versions of the four nucleotides). These allow the reaction to proceed one base at a time and therefore the number of cycles matches the read length. Laser excitation produces fluorescent signals that are recorded for nucleotide identification. Accurate base calling depends on the signal intensity produced by the cluster of clonal DNA.

1.3.2 RNA analysis by deep sequencing

RNA sequencing (RNA-Seq) methods are used for in depth transcriptome analysis through cDNA sequencing at massive scale (Ozsolak and Milos, 2011). In short, the main idea is that a collection of RNA molecules is converted to a library of cDNA fragments which are subsequently sequenced in a high-throughput manner to obtain sequencing reads.

Prior to the sequencing, RNA is isolated from the biological samples of interest. The ribonuclease enzymes (RNases) presence in cells can rapidly degrade RNA and significantly hinder the procedure. For this reason the RNA extraction equipment should be cleaned meticulously and treated with RNase-destroying chemicals. Different methods exist for extracting RNA from samples such as the phenol-chloroform based and the silica-gel based column procedures (Sultan *et al.*, 2014). If the size of the RNA molecules is quite large these need to be randomly fragmented by employing either RNA or cDNA fragmentation methods (Wang *et al.*, 2009).

Frequently the tissue of interest is preserved using a method called formalin fixed paraffin embedded (FFPE) processing, where tissue samples are placed in formalin and subsequently embedded in paraffin (Masuda *et al.*, 1999). While the fixation process preserves the tissue, these steps can lead to severe degradation and chemical modifications of the RNA. This results in highly variable and typically poor quality of the RNA extracted from samples of interest (Roberts *et al.*, 2009), characterised by shorter lengths. Additionally, compared to using the same mass of fresh tissue, the yields obtained are lower (Zhao *et al.*, 2014).

This quality of RNA in terms of degradation is indicated by a measure called RIN (RNA Integrity Number) (Schroeder *et al.*, 2006), produced by an algorithm that can detect presence of degradation products. The RIN score ranges from 1 to 10, where level 10 denotes completely intact RNA. Therefore the lower the score, the lower the quality of the RNA (Schroeder *et al.*, 2006). In general, RIN values greater than 7 are considered good quality.

Following RNA extraction and reverse transcription into a cDNA library, there may be different choices for library preparation which can affect the results. Highly abundant ribosomal RNA (rRNA) that constitutes the majority of total RNA in cells as it is required for protein synthesis (Giannoukos *et al.*, 2012), needs to be removed from total RNA before sequencing when the goal of the analysis is mRNA or gene detection. The standard Illumina protocol addresses this by first isolating total RNA and then selecting messanger RNA with a poly(A) - polyadenylated - purification step. A caveat is that poly(A) purification of degraded RNA may pull out only the most 3' segments of the RNA population, due to the 5' sequence becoming detached from the poly(A) tail (Zhao *et al.*, 2014). Hence, the poly(A) method is not the optimal approach when working with FFPE samples. An alternative approach is rRNA depletion or ribodepletion where the man idea is that rRNA is depleted while preserving the small fraction of messanger RNA (mRNA).

1.4 Clinical diagnostics

Human pathogens can either be viruses, bacteria, parasites or fungi. The traditional process of characterizing infection-causing pathogens in clinical speciments is done through potentially difficult or time consuming techniques. Examples include microscopy and/or cell culture for investigating the microbial composition of a sample, identifying colonies and producing sufficient mass of microorganisms for subsequent use (Didelot *et al.*, 2012). Alternatives include PCR or Sanger sequencing, both introduced in previous section 1.2.2 or DNA microarrays with probes that hybridize known sequences (Yozwiak *et al.*, 2012). Viral detection is frequently laced with additional difficulties as discussed in the following section.

1.4.1 Virological diagnostics and Koch's postulates

Viral pathogens were traditionally detected on cultured cell that exhibited cytopathic effects or plaques or alternatively by antibody neutralization tests (Bibby, 2013). However many viruses cannot be cultured in laboratory conditions while the antibody neutralization tests depend on the availability of quality antiserum. PCR assays are considered the gold standard for diagnostic virology as they are very sensitive, quantitative and inexpensive and can detect unculturable or nonisolated viruses (Lipkin and Hornig, 2015). However PCR relies heavily on prior information on the target viruses and thus, clinical diagnoses of viral infections may require wide arrays of PCRs targeting different viruses. This does not necesarily guarantee a successful outcome, as demonstrated in cases of rare or novel pathogens.

Importantly, the detection of a microorganism in a clinical specimen is only the first step in establishing a causal relationship. Kochs postulates, proposed by Robert Koch towards the end of the 19th century, attempted to establish rigorous criteria and provide guidelines for defining a causative relationship between microorganism and disease. In summary these are the presence of the agent in every case of a disease, specificity for that disease (the agent occurs in no other disease as a nonpathogen) and finally, the capacity to cause the same disease in hosts after repeated propagation in culture (reviewed in (Fredricks and Relman, 1996)). The postulates were soon understood to be non suitable for viruses and a few decades later they were revised (Rivers, 1937). Problems include the fact that several viruses do not cause the disease in all infected individuals. An known example is poliovirus which causes paralysis in about 1% of those infected with the majority of cases either subclinical or non paralytic. Additionally, infection with the same virus may lead to different diseases demonstrated by differences in immunocompotent and immunodefecient/immunosuppressed individuals. Furthermore infection with different viruses may result in similar disease signs and symptoms. Finally, there are viruses that do not replicate in cell culture, or for which a suitable animal model has not been identified.

The postulates were modified to reflect the introduction of culture-independent molecular methods, and were called molecular Koch's postulates (Fredricks and Relman, 1996). There are as follows: firstly, the pathogen is consistently associated with the disease, i.e nucleic acid sequences from the pathogen are present in most cases of the disease. Secondly, the hosts without the disease have either no sequences or smaller numbers of the nucleid acid sequences of the pathogen. Additionally, disease resolution results in decresed numbers of pathogenic sequences. If the sequence copy number correlates with the disease severity the sequence-disease association is more likely to be a causal relationship. Furthermore, the sequence-inferred nature of the microorganism is consistent with the group of organisms it belongs to. Finally, there is specific in situ hybridization of genomic sequence to the areas of tissue pathology and the results providing the evidence for causation are reproducible.

The Koch postulates of causation may be modified and adapted in different times, given changes in technology and disease knowledge (Falkow, 2004). However, even the newer adapted versions cannot be satisfied in all instances despite their relevance to the molecular era (Lipkin, 2008). Infection patterns vary along with factors such as genetic susceptibility, age, nutrition or previous exposure to other agents. In some cases, such as several acute infectious diseases, the responsible microorganism replicates in the tissue of interest, can be readily identified with traditional methods, there is evidence of an adaptive immune response as well as evident morphological changes consistent with infection. However when classical hallmarks of infection are not present, the pathogenesis mechanism is not direct, the microorganism has latent effects or requires cofactors such as coinfection, confirming causation is more challenging. In such cases, the strength of the epidemiological association in the patients needs to be statistically assessed.

Establishing the causal relationship between a virus and a disease is therefore an important but difficult issue that may require multiple separate studies to be ultimately resolved (Fredricks and Relman, 1996). In the most difficult of scenarios this may be obtained only after a specific intervention such as a drug or a vaccine has been shown to prevent the disease (Lipkin, 2009).

1.4.2 Metagenomics

The process of characterizing viruses in clinical samples is being revolutionized by advances in high throughput sequencing. The traditional methods typically focus on identifying a single pathogen at a time and may fail to detect the infectious agent in a significant percentage of cases in some infectious diseases (Tunkel *et al.*, 2008). High throughput sequencing driven methodologies hold the promise of a largely unbiased approach in species detection and of unexpected discoveries, as well as relatively rapid turnaround time (Quail *et al.*, 2012). As a result, researchers have adopted the technology for detecting and characterising either viral or bacterial pathogens responsible for acute and chronic illnesses of unknown origin in isolated cases (McMullan *et al.*, 2012; Wilson *et al.*, 2014; Brown *et al.*, 2015) as well as for disease outbreaks (Rohde *et al.*, 2011; Frank *et al.*, 2011; Chin *et al.*, 2011; Greninger *et al.*, 2010).

Virologists quickly adopted high throughput sequencing for identifying viruses as many viruses cannot be cultured and lack a universal conserved genetic element shared between viral genomes. Especially pertinent to viral infections is the emergence of novel emerging strains which prove additionally challenging for conventional techniques (Brown *et al.*, 2015). Identifying correctly the viruses involved in an illness is crucial for avoiding misdiagnoses that may lead to improper clinical treatment - such as the administration of antibiotics - and which negatively affects survival or transmission rates. The detection and response to viral pathogen outbreaks (Assiri *et al.*, 2013; Cotten *et al.*, 2014; Matranga *et al.*, 2014) is another application of metagenomics. This approach has been successfully used in influenza outbreaks to determine viral subtype (Kuroda *et al.*, 2010; Greninger *et al.*, 2010; Deng *et al.*, 2011). Viral metagenomics also offer the ability to identify coinfections (Yang *et al.*, 2011). Another attractive feature of deep sequencing is the ability to detect variants at low frequencies. This is useful for identifying drug resistant mutations or transmission patterns of the viruses and for evaluating the impact of minority variants on treatment efficacy (Quiñones Mateu *et al.*, 2014). Finally, a number of viruses that have not been associated with human diseases have been detected in healthy human hosts, establishing the existence of a normal human virome (Lecuit and Eloit, 2013).

It is important to note that despite the promise metagenomic identification of viral pathogens offers, there are a number of challenges. The most obvious one is that the probability of inadvertent microbial contamination is not negligible. Microbial contamination can take place during sample handling in the laboratory or from the use of contaminated laboratory reagents or nucleic acid extraction kits. The contaminants may be either bacteria (Salter *et al.*, 2014) or viruses (Naccache *et al.*, 2013; Rosseel *et al.*, 2014). To address this, the same extraction and deep sequencing methods should be applied to both clinical case samples and suitable negative controls, such as blank extractions (Salter *et al.*, 2014).

Inferring causation from metagenomic findings needs to be further supported by methods that depend upon viral particle isolation, such as protein expression, viral replication and reactivity to anti-serum in affected tissues (Bibby, 2013). In general, supportive clinical, epidemiologic and serological data are critical in confirming associations of candidate novel agents with disease. These strategies have been previously used to conclusively refute the putative association of the retrovirus XMRV (xenotropic murine leukemia virus) with chronic fatigue syndrome or prostate cancer and proved that XMRV originated as a mouse cell line-derived laboratory contaminant (Hué *et al.*, 2010; Knox *et al.*, 2011; Simmons *et al.*, 2011).

Finally, a limiting factor of the full potential of deep sequencing is the analysis and interpretation of metagenomic data. The choice of analytic method depends on the aim of the study as well as the computing resources available to the researchers. The use of specific analytic methods actively affects the answer to the question we ask. The methods and the associated limitations and challenges are discussed in the following section.

1.5 Community profiling in metagenomics

Community profiling of a metagenomic mixture is defined as the identification and quantification of the present species in the specific sample. This has proved to be a challenging problem to solve and it raises complex computational issues. Part of the difficulty stems from the read length limitation of existing deep sequencing technologies, an issue compounded by the extensive level of homology across viral and bacterial species. Another complication is the divergence of the microbial sequences from the publicly available references. As a consequence, the assignment of a sequencing read to a database organism is often unclear. Lastly, the number of reads originating from a disease causing pathogen can be low (Barzon *et al.*, 2013), underlying the need for highly sensitive methods. The pathogen contribution to the mixture depends on the biological context, the timing of sample extraction and the type of pathogen considered. Therefore, highly sensitive computational approaches are required.

1.5.1 Related work

A first approach to the problem is read classification, that is the assignment of a given sequencing read to a species. Several tools have been developed and these belong to two broadly defined classes: composition-based and similarity-based approaches. That means that methods generally use the following information sources: sequence composition or sequence identity to reference databases, with hybrid methods using both. A third related approach is to phylogenetically analyze metagenomes by subsetting to core genes that are expected to follow the same evolutionary path and are present in a large proportion of microorganisms. The core genes represent only a small proportion of a metagenome (Matsen, 2015) and thus, the task is not entirely equivalent to read classification. The portion of the remaining metagenome can be taxonomically classified using either similarity or composition based methods.

The read classification based on sequence composition relies on the intrinsic features of the reads, such as CG content or oligonucleotide distributions (Deschavanne *et al.*, 1999; Bentley and Parkhill, 2004). Methods include TETRA (Teeling *et al.*, 2004), PhyloPythia (McHardy *et al.*, 2007), Phymm (Brady and Salzberg, 2009) and LikelyBin (Kislyuk *et al.*, 2009).These tend to focus on major classes in a dataset and may not perform well on low-abundance populations (Kunin *et al.*, 2008). Additionally, results are usually reliable only for longer reads or assembled contigs, usually at least 1,000bp and generally are less accurate compared to similarity based approaches (Dröge and McHardy, 2012). For this reason these methods are not optimal when the goal is to detect a candidate disease agent that is only supported by a low number of short reads in the data.

A tool that leverages phylogenetic analysis of metagenomic sequence data is PhyloSift (Darling *et al.*, 2014). PlyloSift places short sequencing reads or assembled contigs onto a phylogeny of core reference genes which include viral gene families. PhyloSift relies on a relatively small set of widely conserved marker genes so there is little informative variation at high taxonomic resolution. PhyloSift is not a taxonomic classification method but rather its interest is in providing the phylogenetic framework for the deduced informations and is more intended for quantification of abundances of relatively large clades.

Similarity based methods, using similarity search algorithms such as BLAST (Altschul *et al.*, 1990), are considered the most sensitive methods for read classification (Brady and Salzberg, 2009). One of the most popular tools using the output of a similarity search algorithm is MEGAN (Huson *et al.*, 2007). MEGAN addresses ambiguous matches by assigning reads that have multiple possible assignments to several species to the taxonomic group containing all these species, or else their lowest common ancestor (LCA). This approach is accurate on a higher taxonomic level. However, it is lacking a formal solution to resolving ambiguous matches. Kraken (Wood and Salzberg, 2014) couples the LCA approach with exact *k*-mer (words of *k* nucleotides) matching. Its use of *k*-mers may result in low specificity, performing better on genus-level classification. Additionally, a single value for Kraken's *k*-mer might not work equally well for viral and bacterial genome due to the viral higher mutation rate, hampering its performance in samples where both co-exist. CLARK (Ounit *et al.*, 2015) is another recent *k*-mer based assignment tool. MetaPhlAn2 (Segata *et al.*, 2012) uses a reference database consisting of clade-specific defining genes, as opposed to PhyloSift's universally conserved genes. Similarly it cannot classify non-marker reads, additionally it was developed primarily for profiling metagenomic samples where the coverage of many microbes is reasonably high, therefore is suboptimal for detecting traces of a microbe. MetaPhlAn2 does not try to classify each read but it rather focuses on estimating relative abundances.

An obvious general limitiation of similarity based methods is their reliance on the content and completeness of public reference databases. Public databases hardly represent the real biological diversity, especially pertinent for the viruses that are mostly undiscovered (Fancello *et al.*, 2012). Additionally their content is biased towards cultivable organisms and human pathogens (McHardy and Rigoutsos, 2007). Therefore reads from novel microorganisms that are sufficiently divergent from known species will either be misclassified or unclassified. The resulting classification will differ depending on the choice of the database. It is possible that using the same database source but a different updated version with improved annotations and new additions may produce different results. Informed database selection can limit the presence of false positives, however more important is to bear in mind the database limitations and biases and interpret findings with care. In the event of an exciting but unlikely finding, researchers need to proceed with caution and apply real life Bayesian reasoning. Different sanity checks, bioinformatic ones but also in the form of new experiments are crucial for accepting a finding.

The type of the database is also important when the goal is to uncover viral sequences in the sequencing data. Viruses in general are characterized by high genetic diversity and divergence (Fancello *et al.*, 2012) therefore we are more frequently interested in recovering remote similarities. A higher level of conservation is expected at the protein level compared to nucleotides, therefore protein databases that allow peptide rather than nucleotide similarity searches, are suggested (Kunin *et al.*, 2008).

Another weakness of the similarity based methods is that a long tail of species, each supported only by few reads can appear in the results. This is due to two factors: first, the classification is decided one read at a time, in contrast to considering all reads simultaneously. Second, while the more complex read interpretation fits the data better, it can lead us to infer overcomplicated profiles, i.e assuming thousands of species to be the origin of the sequencing reads. However any proposed explanation should not be more complicated than necessary; consideration of the plausibility of the models (profiles) needs to be part of the inference. Bayesian methods automatically incorporate this "Occam's razor" in the form of priors for the considered models (Mackay, 1992). Even when researchers do not explicitly think in terms of prior probabilities for their hypotheses, when they consider simple hypotheses before complex ones, they are in fact intuitively doing exactly that.

1.5.2 Similarity based - mixture model approach

Methods designed to infer the set of present species as well as estimate their relative proportions, incorporate knowledge from all reads to assign each individual read to a species. From a statistical standpoint, this identification and quantification question can be thought of as an application of finite mixture models. Mixture models have been applied in the metagenomics context in frequentist and Bayesian settings. GRAMMy (Xia *et al.*, 2011) formulates the problem as a finite mixture model, using the Expectation-Maximization (EM) algorithm to estimate the relative genome abundances. However it cannot work with BLASTx output which is better suited for viral discovery for reasons explained in the previous section. Pathoscope (Francis *et al.*, 2013) refines this process by penalizing reads with ambiguous matches in the presence of reads with unique matches and enforcing parsimony within a Bayesian context.

Fitting a mixture model is useful for the species relative abundance estimation as well as for the read to species assignment. A related but distinct question concerns the set of species which should be included in the mixture model. This question is closely related to the biological question of asking what species are present in the mixture. Including all species flagged as potential matches by the read annotation can introduce a large number of species, often in the low thousands. Mixture models will then identify a large number of species at low levels. This interpretation is appropriate in some applications. In many other cases, the expectation is that the underlying species set should be parsimonious and that some divergence with database species or sequencing errors can explain a large fraction of the non matching reads.

Pathoscope is the only other community profiling approach to use a Bayesian statistical framework and thus shares a degree of similarity to the method I propose and developed for this thesis, called metaMix and introduced below in section 1.7. Pathoscope was not published until the end of 2013, towards the end of metaMix development. metaMix is significantly different to both GRAMMy and Pathoscope in employing a parallel MCMC approach to explore the state-space of the candidate organisms by comparing different combinations of species based on their posterior probabilities.

1.6 Bioinformatics processing

Before we can describe the proposed method for community profiling, we go through the bioinformatics analysis steps that takes us from the raw sequences to the type of data metaMix can use. We also briefly talk about the questions that may become relevant when we have the community profile of a clinical sample.

1.6.1 Prior to community profiling

Prior to applying the community profiling approach using Bayesian mixture models, several steps are required to process the short read sequence data. The pipeline uses publicly available bioinformatics tools for each preprocessing step.

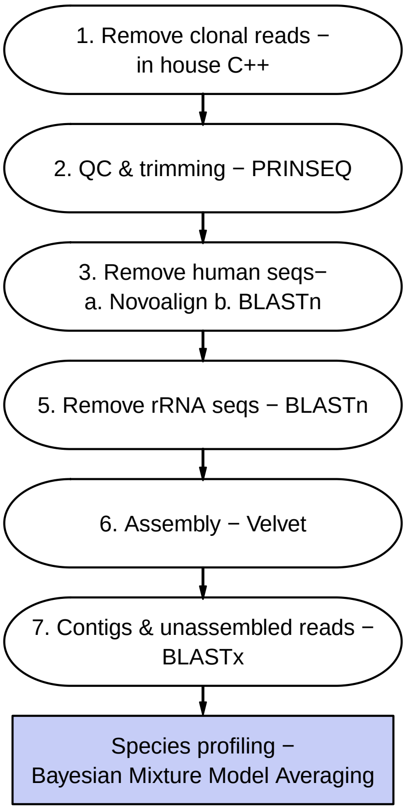

1. Remove clonal reads − in house C++
2. QC & trimming − PRINSEQ
3. Remove human seqs− a. Novoalign b. BLASTn
5. Remove rRNA seqs − BLASTn
6. Assembly − Velvet
7. Contigs & unassembled reads − BLASTx
Species profiling − Bayesian Mixture Model Averaging

Figure 1.5: Bioinformatics pipeline steps prior to species identification. The step of removing the rRNA sequences is applicable only when the aim is viral discovery.

1.6.1.1 Removal of clonal reads

The very first step is the removal of clonal reads using an in house C++ script, implemented by Vincent Plagnol. Duplicate sequences are a common issue in Illumina sequencing (Kozarewa *et al.*, 2009). Standard Illumina library preparation involves

PCR amplification before the sample is loaded into the flowcell and PCR can cause clonal artifacts. Duplicate reads are typically defined post-alignment to a reference genome as reads that begin and end at exactly the same start and end coordinates. Collapsing clonal reads reduces the number of reads, hence facilitating all downstream analyses. Furthermore we infer the relative abundances based on the read counts, so such duplicates could lead to overestimate the species abundance.

We cannot use position information to detect identical reads as we are interested in all reads and not reads matching a specific organism such as the host. Instead, the sequence information of paired-reads is used. If identical, the pairs are being collapsed keeping the one with the best average pair quality. We do not require identity across the whole length of the read, instead the first n nucleotides are compared, as it is typical for Illumina reads to have low quality ends. The number n depends on the read length, but we typically use $\sim 80\%$ of the read length as signature.

1.6.1.2 Quality control

The quality of the 3' end of a sequencing read can be low due to phasing artifacts (Metzker, 2010). This term describes failures in nucleotide incorporation or in block removal, or incorporation of more than one nucleotide in a particular cycle. This results in uneven strand lengths, introducing shorter and longer strands within the same cluster. In turn, this reduces the purity of the signal output and as the variation in strand lengths increases with every cycle, the precision of base calling drops (Erlich and Mitra, 2008). We use PRINSEQ (Schmieder and Edwards, 2011) for read-based quality control, removing low quality and complexity reads and performing 3'end trimming.

1.6.1.3 Host filtering

For metagenomic analysis of human samples, reads originating from the human host are usually not relevant for our research question. We therefore remove human host reads, using a two-step approach to limit computation time: initially a short read aligner (`novoalign`, www.novocraft.com), followed by BLASTn.

It is important to first define homology and similarity in the context of sequence analysis before we can describe the basis of BLAST and `novoalign`. Sequence analysis aims to find important sequence similarities that would allow one to infer homology. Homology is defined as common origin which means that sequences are homologous if they share a common evolutionary ancestry. Similarity on the other hand is an observed quantity and we usually talk about percent identity of two sequences. Homology can be inferred by sequence but also structural and functional similarity. When two sequences or structures share more similarity than would be expected by chance, we can infer that the two sequences did not arise independently but rather from a common ancestor.

There are different algorithms for pairwise sequence alignment, that is the comparison of two sequences to establish regions of residue similarity. The Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) provides a method for finding the optimal alignment over the entire length of two sequences. This method maximizes the number of amino acid matches, assign scores to mutations, insertions and deletions and computes an alignment of two sequences that corresponds to the least costly set of such changes. Needleman-Wunsch is a global alignment technique and therefore it cannot be used to find local regions of high similarity. Local sequence alignment is frequently more useful because there is greater variation (insertions, deletions, mutations) towards the protein sequences ends which are less conserved. The Smith-Waterman algorithm (Smith and Waterman, 1981) performs local alignment and finds the local region of highest similarity between two proteins without having to align their ends that may be highly different. Local alignment is useful for finding sequences that have low similarity and different lengths.

Both the Needleman-Wunsch and the Smith-Waterman algorithms are optimal sequence alignment methods which find the highest scoring alignment for any pair of protein sequences. These algorithms tend to be slow and performing a sequence alignment in reasonable time is difficult as reference databases increase in size. BLAST (Altschul *et al.*, 1990) is a heuristic algorithm based on Smith-Waterman and it was designed to offer a balance between sensitivity and speed. BLAST first finds all words of a specific length that exist in the query protein (nucleotide) sequence. BLAST then finds all the closely related words with conservative substitutions introduced using a substitution matrix, that is a matrix of similarity scores for all possible pairs of residues. Local alignments are extended in both directions until the gaps and the mismatches result in the score of the alignment to drop more than a prespecified amount.

Unlike BLAST (Altschul *et al.*, 1990), whose original purpose is finding homologous sequences to query protein sequences by searching though large databases, short-read aligners are generally used for the alignment of DNA sequence from the species of interest to the reference genome assembly of that species. This translates into expecting mismatches to be driven by the species polymorphism rate and the technology error rate rather than by evolutionary substitutions, hence the big speed gain in processing million short reads (Flicek, 2009). Different short read aligners are based on the Burrows-Wheeler transform (BWT) of the reference genome. BWT is an algorithm that can yield a more compressible dataset (Burrows and Wheeler, 1994) and only works well with references that are larger than several thousand characters. This data indexing technique therefore maintains a relatively small memory footprint when searching through a given data block. Examples of popular aligners that use BWT include BWA (Li and Durbin, 2009), SOAP2 (Li *et al.*, 2009b) and Bowtie2 (Langmead and Salzberg, 2012).

Novoalign, the short aligner chosen for this step builds the genome index by dividing the reference sequences into overlapping kmers. The alignment process first finds alignment locations in the indexed reference sequence that are possible sources of the read. The alignment locations are then scored using the Needleman-Wunsch algorithm with affine gap penalties and position specific scoring. The latter is derived from the read base qualities and the ambiguous codes in the reference sequence. Novoalign was chosen for this step due to greater experience with this tool and different aligners could be used to perform this step without major differences in the final results.

The next step is only applicable when the focus is on virus discovery using transcriptome reads. We remove ribosomal RNA sequences to avoid false positive alignments to viruses that share sequence similarity with human ribosomal RNA. We use BLASTn against the Silva rRNA database (`http://www.arb-silva.de/`).

1.6.1.4 *de novo* assembly

The remaining reads are assembled into contigs using the Velvet short read assembler (Zerbino and Birney, 2008). Assembly is the process of combining short read fragments into contiguous stretches of DNA called contigs. In the era of dealing with long Sanger reads and single genomes, assembly tools used the overlap graph (Myers, 1995). The main idea of the overlap graph is straightforward and intuitive: each read is regarded as a node and two reads presenting a clean overlap are connected by a bidirected edge.

In the deep sequencing context, the huge number of reads make the overlap graph extremely large, requiring large computational resources.

De Bruijn graphs (Pevzner *et al.*, 2001) are the basis of a very different approach. The fundamental difference is that graphs are not centered around reads but around k-mers, words of k nucleotides. A De Bruijn graph, monitors overlaps of $k-1$ length between these k-mers instead of overlaps between the actual reads. Reads are mapped as paths through the graph, going from one word to the next in a determined order. High redundancy is naturally handled by the graph without affecting the number of nodes. Searches for overlaps are simplified, as overlapping reads are mapped onto the same arcs and can easily be followed simultaneously (Zerbino and Birney, 2008). De Bruijn assemblers include Velvet (Zerbino and Birney, 2008), ALLPATHS (Butler *et al.*, 2008), SOAPdenovo (Li *et al.*, 2010). We chose to work with Velvet, one of the first methods to be developed and still very popular..

These assembly approaches assume uniform coverage and linear genome sequence. These are normal assumptions for a single genome assembly which no longer hold in the metagenomics scenario, where the coverage fluctuates both between organisms and in the case of metatranscriptomics, expressed genes of an organism. Therefore, the coverage can no longer be used to determine the uniqueness of regions or to isolate erroneous sequence. Other likely problems are that highly conserved sequences shared between different species can cause chimeric contigs and sequences of highly abundant species could be misidentified as repeats.

As part of the pipeline a conservative assembly is attempted in order to avoid chimeric contigs. A Velvet tuning parameter is the user defined k-mer length that specifies the extent of overlap required to assemble read pairs. Short k-mers work best with the low abundance organisms, while long k-mers with the highly abundance ones. The shorter the k-mer the greater the chance of spurious overlaps, hence we choose relatively high k-mer length, in order to avoid chimeric contigs.

The approach settled for is not necessarily the one that produces the greater number of contigs or captures both high and low abundance species. The first year of my project I worked extensively on optimizing the assembly step, based on the hypothesis that all downstream inference would be based only on contigs. This work resulted in the realisation that unassembled reads provide crucial information for rare species detection. Some of the work that lead to this choice is outlined in the Appendix A. Following the assembly step, we record for each contig the number of reads required to construct it. We use this information at the stage of species abundance estimation.

1.6.1.5 Annotation of reads and contigs

Finally, for each contig and unassembled read we record the potentially originating species, using the nucleotide to protein homology matching tool BLASTx. We use BLASTx due to the higher level of conservation expected at the protein level compared to nucleotides. This choice is guided by our focus on viral pathogens - viruses having high genetic diversity and divergence (Fancello *et al.*, 2012).

We use a custom reference database that combines viral, bacterial, human and mouse RefSeq proteins for annotation of the reads and contigs. All viruses are used (`ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.1.protein.faa.gz`) as well as all the bacteria of the human microbiome, according to `ftp://ftp.ncbi.nih.gov/genomes/HUMAN_MICROBIOM/Bacteria/all.faa.tar.gz`. The reference database are clearly not complete representations of the biological diversity. However for the purposes of this work and the type of analysis we perform, the complete collection of viruses - human and non human - should be able to capture even novel viruses that are not greatly divergent from all known viruses to date. The human proteins will help classify correctly human sequences that were not filtered out during the two-step host removal and mouse sequences to capture potential laboratory contamination. In the type of data I have analysed for this thesis pathogenic bacterial infections are not very likely, however the bacteria of the human microbiome are included to help capture bacterial contaminants.

If taxonomic information is not included in the BLAST output, we obtain it by using the NCBI taxonomy files that map proteins to taxons. For simplicity we subsequently drop the protein information and only keep a record of mismatches between the read and the species. If a read matches multiple proteins from the same species, we keep only the best match. This step generates a sparse dissimilarity matrix between the read sequences and the protein sequences, as in Table 1.1, with species as columns, reads and contigs as rows.

Table 1.1: Input data for community profiling - sparse dissimilarity matrix

	species$_1$	species$_2$	...	species$_K$
read$_1$	m$_{11}$	m$_{12}$	...	m$_{1K}$
read$_2$	m$_{21}$	m$_{22}$	...	m$_{2K}$
		...		
read$_N$	m$_{N1}$	m$_{N2}$	...	m$_{NK}$

where m_{11} is the number of mismatches between the sequence of read$_1$ and the sequence of the "best match" protein from species$_1$.

The pipeline can be found online http://github.com/smorfopoulou/clinicalDiagnostics_pipeline as it has been shared with UCL Genomics (UCL sequencing services facility) to be used for the analysis of any generated metagenomics datasets from human clinical samples.

1.6.2 Post community profiling

After the community profiling step has taken place, one may be able to answer further research questions regarding a specific species of interest, for example a potential pathogen detected in a clinical sample. This is conditional on having a sufficient number of reads originating from the pathogen so that a *de novo assembly* can be performed to recover its full genome sequence. Such a question could be, how does the detected organism compare to different strains of the same species? We may answer this by inferring a phylogenetic tree. Phylogenetic trees generally represent the evolutionary relationships within a group of organisms (Delsuc *et al.*, 2005).

A phylogenetic tree contains nodes connected by branches. There are external nodes called leaves or tips, which correspond to the actual sequences we are interested in estimating a phylogeny for. The internal nodes represent the putative common ancestors of the descendants (the tips). The branch lengths indicate the amount of evolutionary time along the branches and are expressed in units of expected number of substitutions per site. Phylogenetic trees are inferred from the sequence data with reconstruction methods falling in two general categories: distance-based and character-based methods (Yang and Rannala, 2012). In the first category a distance matrix is used for the tree reconstruction, with a distance calculated for all pairs of sequences. The neighbour joining algorithm (NJ) (Saitou and Nei, 1987) is a popular distance-based method, its main advantage being the computational efficiency and resulting speed. The method performs well when the sequences for which we want to estimate the phylogeny are not very divergent. However sacrificing information by using distances instead of the full sequence information hinders the reliable estimation of pairwise distances for divergent sequences (Holder and Lewis, 2003).

Maximum likelihood (ML) (Felsenstein, 1981; Yang, 1993), maximum parsimony (MP) (Day, 1987) and Bayesian inference methods (Huelsenbeck *et al.*, 2001; Holder and Lewis, 2003) are character-based methods. These methods simultaneously compare all sequences considering a single position in the multiple sequence alignment at a time in order to calculate a tree score. This is the log-likelihood value for ML, the minimum number of changes for MP and the posterior probability for the Bayesian methods. A comparison of all potential trees reveals the tree with the best score, however in practice due to the great number of trees this exhaustive search is not possible. Therefore a starting point is created with the help of a fast algorithm that produces an approximate tree, followed by local rearrangements to improve the tree score. These methods are naturally much slower compared to NJ however the added accuracy justifies their use and popularity.

A second interesting question is whether variable sites exist between our detected genome and a genome of interest, typically named as the reference genome. The identification process is called variant calling and it is based on the observed nucleotide counts at a single sequence position. Read coverage, base qualities, variant frequencies and strand bias are of great importance during variant identification in order to differentitate real variants from errors (Koboldt *et al.*, 2012; McElroy *et al.*, 2013, 2014). A positional error model can be incorporated in the detection process (Flaherty *et al.*, 2012; Wilm *et al.*, 2012). The reference genome may be a publicly available genome sequence or the sequence of a currently circulating pathogen strain. When the goal is to identify minority variants (variants of low frequency) within the pathogen population in the clinical sample, the reference genome becomes the *de novo* assembled sequence generated by the short read data.

Only in one case out of all the clinical samples analysed during the PhD project, discussed in chapter 5 there were enough reads to recover the full pathogen genome from a *de novo* assembly. The phylogenetic tree was reconstructed using PhyML (Guindon *et al.*, 2010), a maximum likelihood approach. Variants were identified using

SAMtools (Li *et al.*, 2009a; Li, 2011).

1.7 Objectives and proposed method

We previously discussed some of the limitations of the current classification methods limitations, especially when the setting is potential pathogen discovery where low abundance organisms are of interest. A few of the methods are designed to work with viruses, including sets of viral marker genes or offering support for BLASTx results. However the most pronounced limitation is the high number of false positives in the results. This drawback may cause users to focus on the higher abundance organisms in the results, as the lower abundance organisms may well be false positives. Depending on the research question, this can be a legitimate approach, however it may become problematic for scenarios where detection of even minute amounts of a potential viral pathogen in the sample is of interest. There are statistical tools we can use in order to address this issue in a formal and elegant way. We propose to classify a single read by borrowing information from the whole of a dataset and to consider the plausibility of different community profiles. The main objective of this thesis is thus to develop an open-source, sensitive, specific and accurate similarity based community profiling method at a high taxonomic resolution, employing these two ideas.

Similarity methods use the output of tools that perform similarity searches and record this information for pairs of query and subject sequences. Similarly, metaMix works with the output of BLAST which can be rewritten as a sparse matrix that records the mismatches between each sequence in the dataset and each sequence in the reference database, as presented in Table 1.1.

metaMix employes a parallel MCMC approach to explore the state-space of the candidate organisms by comparing different combinations of species based on their posterior probabilities. We can decompose the proposed method into two levels of inference, one nested within the other. First we fit each model that represents a different profile, that is a different combination of species, to the data. This first level corresponds to the estimation of the model's parameters given the model and the data. Prior information or prior belief about the relative plausibility of the competing models can be incorporated in the inference. Therefore second, we perform model comparison in the light of the data, assigning suitable priors to the alternative models in order to penalise overly complex models.

The proposed method for the community profiling problem requires the exploration of the candidate organisms state-space. The main challenge is computational; even with a relatively small number of species to consider, the number of subsets of this space that could explain the mixture grows exponentially. Efficient computational strategies are required to make this problem tractable, so that the inference can be achieved for modern scale metagenomics datasets. Our strategy is based on Parallel Tempering, a Monte Carlo Markov Chain technique, using parallel computing to speed up the inference. Within the Bayesian framework readily interpretable probabilities such as the posterior probabilities of species sets can be used to quantify the support for a species in the mixture.

I implemented metaMix in an R package, available on CRAN (`http://cran.r-project.org/web/packages/metaMix`) and described in detail in chapters 3 and 4. metaMix produces posterior probabilities for various models as well as the relative abundances under each model. Its performance and potential is demonstrated using clinical samples, discussed in chapters 5 and 6 as well as benchmark metagenomic datasets in chapter 4.

1.8 Thesis outline

The outline of the remaining thesis is as follows. Chapter **2** provides an introduction to the statistical concepts that are central to this work: basics of Bayesian inference, Monte Carlo methods, basic MCMC techniques, introduction to Parallel Tempering, theory of Mixture Models as well as the missing data formulation. Chapter **3** connects the concepts introduced in chapter **2** to the metagenomic community profiling problem. with a description of how they are used to build the novel methodological framework. Chapter **4** illustrates features of the metaMix R package using a toy dataset. metaMix performance is assessed using benchmark metagenomic datasets. Chapter **5** demonstrates the successful use of metaMix towards pathogen identification in two clinical samples from patients with undiagnosed encephalitis. Chapter **6** describes the project that was the original motivation for developing metaMix. This work aimed to investigate whether there viruses are triggering the onset of Type I Diabetes. Chapter **7** provides the conclusions and explores possible directions for future research.

Chapter 2

Bayesian methods and Monte Carlo strategies

In this chapter we introduce the statistical concepts that are central to this thesis. These will appear multiple times and are fundamental for understanding the proposed methodology for performing community profiling based on metagenomic data. We start by discussing Bayesian methods in general, followed by an introduction to Markov Chain Monte Carlo methods and Parallel Tempering MCMC, a discussion on Bayesian model choice and model averaging and finally, a brief overview of Finite Mixture Models theory.

2.1 Bayesian methods

Bayesian methods (Jeffreys, 1961; Gull, 1988) were initially introduced by Bayes and Laplace in the 18th century and subsequently developed in the 20th century. Generally in the Bayesian context the parameters θ are treated as random variables. Therefore parameters are described by distributions, representing our belief about the parameters values given the observed data and prior knowledge. According to Bayes theorem, the posterior distribution that describes the probability of the parameters θ given the data can be computed in terms of the data likelihood $P(X|\theta)$ and the prior distribution $\pi(\theta)$. The likelihood of the data is the probability of X conditioned on θ. The prior distribution is our a priori belief about the value of parameter θ before observing the data X. The posterior probability distribution of the parameters θ given the data X is:

$$P(\theta|X) = \frac{P(\theta,X)}{P(X)} = \frac{P(X|\theta)\pi(\theta)}{P(X)}. \tag{2.1}$$

Certain choices for the likelihood and the prior can result in a convenient form for the posterior distribution. In particular, there are cases in which for a given form of the likelihood function, if the prior belongs to a family of distributions, then the posterior also belongs to the same family. This property is known as closure under sampling (Gelman *et al.*, 2003) and we say that the prior and posterior are conjugate distributions. For example, when the likelihood is modeled as a multinomial, the conjugate prior can be a Dirichlet distribution. That means that the posterior will also be a Dirichlet distribution. The main motivation for using conjugate priors is their tractability, however they can be constricting for the user.

The normalising constant $P(X)$:

$$P(X) = \int P(\theta, X) d\theta = \int P(X|\theta)\pi(\theta) d\theta \tag{2.2}$$

is the average of the likelihood over the whole parameter space. This is an important quantity as it represents the evidence for a particular model. $P(X)$ is also called marginal likelihood as we marginalise out θ from the joint distribution $P(\theta, X)$. We will discuss the issue of marginal likelihood estimation in a few sections.

Once the posterior distribution is available, features of θ such as means and variances of the individual θ_i require integrating over the posterior distribution. Therefore Bayesian inference problems can be expressed as the expectation of a function of interest $f(\theta)$, evaluated over the posterior distribution:

$$\mathbb{E}[f(\theta)] = \int \pi(\theta|X) f(\theta) d\theta. \tag{2.3}$$

In simple cases such integrals can be computed analytically, however in most realistic problems computational methods are required. Simulations play a central role in Bayesian analysis due to the relative ease with which samples can often be generated from a probability distribution. Generating values from a probability distribution is straightforward with modern computing techniques using pseudorandom number generators.

2.2 Monte Carlo sampling

The main premise of Monte Carlo sampling is to use random samples from a specified probability distribution to approximate difficult to evaluate integrals. In Bayesian analysis, this generally involves acquiring samples from the posterior distribution. Monte Carlo methods approximate the expectation of a function of interest $f(\theta)$ evaluated over the posterior distribution $\pi(\theta|X)$, or else the integral 2.3, by drawing samples θ_i, $\{i = 1, \cdots, n\}$ independently and randomly from $\pi(\theta|X)$. Additionally, if f is scalar-valued then $\mu = \mathbb{E}[f(\theta)]$ and $s = \sqrt{Var(f(\theta))}$ are also scalars. We take the average as our estimate of μ:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\theta_i). \tag{2.4}$$

With independent samples θ_i the approximation can be made as accurate as desired by increasing the sample size n, as ensured by the law of large numbers (Robert and Casella, 2004). The mean of the estimator $\hat{\mu}$ is:

$$\mathbb{E}[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(\theta_i)] = \mu. \tag{2.5}$$

Therefore the Monte Carlo estimator is unbiased for μ. Additionally the variance of the estimator is:

$$Var(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mu)^2] = \frac{1}{n^2}\mathbb{E}[(\sum_{i=1}^{n} f(\theta_i) - \mathbb{E}[\sum_{i=1}^{n} f(\theta_i)])^2] = \frac{1}{n^2}Var(\sum_{i=1}^{n} f(\theta_i)) = \frac{s^2}{n}. \tag{2.6}$$

Therefore the standard deviation of the Monte Carlo estimator is $\frac{s}{\sqrt{n}}$. It is of importance to note that the dimension of the space for the function f does not appear in the formula. The Central Limit Theorem (CLT) tells us that the Monte Carlo converges at a rate of $\frac{1}{\sqrt{n}}$ (Liu, 2001), the error is normally distributed for large n and the complexity of the computation depends only on the variance of θ. This is the second attractive feature of Monte Carlo which motivates its use in high dimension problems, in contrast to numerical integration methods which suffer from the curse of dimensionality. However it is worth noting that the computation of $f(\theta)$ may also become challenging as when f is complicated to compute then that step will also take time.

The basic Monte Carlo approach is to sample points randomly from the specified probability distribution. An obvious issue is that a lot of effort can be wasted in evaluating random samples located in regions where the function value is almost zero. Therefore the basic Monte Carlo typically suffers from low efficiency. The uncertainty of the estimator can be decreased by increasing n, but this converges very slowly. There are ways to overcome this using variance reduction techniques, such as Importance Sampling.

A second issue relevant to Bayesian analysis is that it is not always easy to draw from a complex posterior distribution directly. In simple Bayesian models, especially if conjugate prior distributions have been assumed this can be straightforward, however in more realistic problems this will not be the case. We might instead generate independent samples from some simpler approximating distribution and then compensate for the use of the wrong distribution. Alternatively, with non-independent samples θ_i, equation (2.4) still holds when the samples are drawn throughout the support of $\pi(\theta|X)$ in the correct proportions. We can achieve this by simulating a Markov chain that converges to the correct distribution.

2.3 Importance Sampling

Importance Sampling (IS) (Liu, 2001) is a method of determining the properties of a distribution by drawing samples from another distribution. More specifically IS is a method for computing expectations using random samples drawn from an approximation to the target distribution. The main idea of IS is to use importance functions instead of the original distributions, in order to focus on regions of importance and not waste simulations. We are interested in evaluating 2.3, given an arbitrary density g that is positive when $\pi(\theta)f(\theta)$ is not zero.

The algorithm steps are the following: we first choose an efficient IS proposal distribution $g(\theta)$, generating n samples from it. We then compute the IS weights:

$$w_i = \frac{\pi(\theta_i)}{g(\theta_i)} \tag{2.7}$$

and we approximate 2.3 by:

$$\hat{I} = \sum_{i=1}^{n} f(\theta_i)w(\theta_i) \tag{2.8}$$

which is the unbiased IS estimator. In equation 2.8 the ratio $\frac{\pi(\theta)}{g(\theta)}$ needs to be known exactly. When this is not the case, we use the following weighted IS estimate:

$$\hat{I} = \frac{\sum_{i=1}^{n} f(\theta_i) w(\theta_i)}{\sum_{i=1}^{n} w(\theta_i)} \tag{2.9}$$

where the ratio $\frac{\pi(\theta_i)}{g(\theta_i)}$ only needs to be known up to a multiplicative constant.

A common problem with IS is that the selection of a good importance sampling distribution can be difficult, with a poor choice resulting to mediocre estimators with infinite variance. This is especially pronounced when the importance weights are small with high probability but very large with a low probability. This can happen if $\pi(\theta)f(\theta)$ has wide tails compared to $g(\theta)$. This difficulty can limit the applicability of the method and for this reason we need to introduce more sophisticated sampling algorithms based on Markov chains.

2.4 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) (Gilks, 1999), (Robert and Casella, 2004) is essentially Monte Carlo integration using Markov Chains. MCMC techniques are often used to solve integration and optimisation problems in high dimensional spaces. As discussed before, integration has a central role in Bayesian statistics and MCMC is a known general approach for providing a solution within a reasonable time, when it is not numerically tractable. Some typically intractable integration problems are involved in making inference for model parameters or making predictions. The main idea for the MCMC methods is that they can construct a Markov chain whose stationary distribution (or else invariant distribution) is the posterior distribution of interest. We briefly introduce the theory underlying MCMC methods, and we then describe the general form of MCMC given by the Metropolis-Hastings algorithm.

2.4.1 Markov Chains

Assume a sequence of random variables $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \cdots\}$ such that given the current state $\theta^{(t)}$ at each time $t \geq 0$, the distribution of the next state $\theta^{(t+1)}$ does not depend further on the history of the chain $\{\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(t-1)}\}$. The distribution of the initial state $\theta^{(0)}$ and the transition kernel $Pr(\theta^{(t+1)}|\theta^{(t)} = \theta)$ define the Markov Chain $\{\theta^{(t)}\}$.

A Markov Chain is said to have stationary distribution $\pi(\theta)$ if:

$$\theta^{(t)} \sim \pi(\theta) \Rightarrow \theta^{(t+1)} \sim \pi(\theta). \tag{2.10}$$

A Markov chain displays detailed balance for the distribution $\pi(\theta)$ if the following holds for any two states $\theta^{(t)}$ and $\theta^{(t+1)}$:

$$\pi(\theta^{(t)})P(\theta^{(t+1)}|\theta^{(t)}) = \pi(\theta^{(t+1)})P(\theta^{(t)}|\theta^{(t+1)}). \tag{2.11}$$

Finally a Markov chain is irreducible, if there is positive probability to move from any state to any other state in a finite number of steps.

2.4.2 The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm (Hastings, 1970), (Chib and Greenberg, 1995) simplifies the task of creating a chain whose stationary distribution corresponds to the posterior distribution. With the MH algorithm at each time t, the next state $\theta^{(t+1)}$ is chosen by first sampling a candidate point θ' from a proposal distribution $q(.|\theta^{(t)})$. The proposal distribution q can have any form and it may also depend on the current point $\theta^{(t)}$. The candidate point θ' is accepted with probability $\alpha(\theta,\theta')$ where:

$$\alpha(\theta,\theta') = min(1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}). \tag{2.12}$$

If the candidate point is accepted, the next state becomes $\theta^{(t+1)} = \theta'$, otherwise $\theta^{(t+1)} = \theta^{(t)}$. The steps of the algorithm are described below in Alg. 2.1.

Algorithm 2.1 Metropolis Hastings algorithm

```
• Initialization $\theta^{(0)}$
• At iteration $t$
    1. Sample a point $\theta'$ from $q(.|\theta^{(t)})$ .
    2. Sample a Uniform(0,1) random variable $U$ .
    3. If $U \leq \alpha(\theta,\theta')$ set $\theta^{(t+1)} = \theta'$ else set $\theta^{(t+1)} = \theta^{(t)}$ .
```

To control for the effect of where in the distribution the chain was initialized, the initial part of the posterior samples is discarded as burn-in. Burn-in samples are the initial samples that are not representative of the stationary distribution. Depending on the context, different fractions of burn-in may be appropriate.

2.5 Parallel Tempering MCMC

In some cases simple MCMC methods are not able to correctly traverse the state space. This is because the models used to analyse the data are often complex leading to poor mixing of the chains. That means that Markov chain simulations may remain in the neighborhood of a single mode for a long period of time. This occurs primarily when different modes are separated by regions of low posterior density. Then it is difficult to move from one mode to the other because jumps to the region between the two modes will be rejected. Such a chain will move between modes only rarely, taking a long time to reach equilibrium.

A potential solution is to perform parallel tempering MCMC (PT MCMC) (Brooks, 1998; Earl and Deem, 2005) which relies on a family of "tempered" distributions, each of which is obtained by varying a temperature parameter T. Each chain simulates from the posterior distribution $\pi(\theta)$ raised to a temperature

$$\{t_1 = \frac{1}{T_1}, t_2 = \frac{1}{T_2}, \ldots, t_{max} = \frac{1}{T_{max}}\}. \tag{2.13}$$

The different temperature levels result in tempered versions of the posterior distribution $\pi(\theta)^{t=1/T}$, where $t \in (0,1]$. When temperature is low and more specifically when $T = 1$, the draws are from the posterior distribution, i.e $\pi(\theta)^{t_1} = \pi(\theta)$. While $\pi(\theta)^{t_j}$ is not too different from $\pi(\theta)^{t_{j+1}}$, the temperature ladder results in a considerable difference between $\pi(\theta)$ and $\pi(\theta)^{t_{max}}$ in that the latter has fewer isolated modes. In practice that means that the posterior distribution at higher temperatures spreads out its mass and becomes flatter, so that it is considerably easier to sample using MCMC techniques. The basis for any subsequent inference is the sample path of the chain with the correct stationary distribution, i.e solely the original posterior distribution with $T = 1$.

Therefore the main idea of PT MCMC is to allow a collection of n synchronised Markov Chains run in parallel and exchange information probabilistically, creating global moves that result in faster mixing. The algorithm is summarized in Alg. 2.2.

Algorithm 2.2 Parallel Tempering MCMC algorithm

1. Initialization of Markov chain (done for all n chains).

2. **Mutation step** at iteration t (done for all n chains).

   - Acceptance probability:

     $$\alpha(\theta,\theta') = min(1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}). \tag{2.14}$$

     $q(\theta|\theta')$ the probability of transitioning from θ' to θ.

   - If the step is accepted, the chain moves to proposed state θ'.

   - If not accepted, the chain's current state becomes the previous state of the chain.

3. **Exchange step** when all chains have advanced a prespecified number of iterations, e.g one iteration.

   - Proposes to swap the value of 2 chains adjacent in terms of T, respective chain values θ_i and θ_j, respective temperatures $t_1 = \frac{1}{T_1}$ & $t_2 = \frac{1}{T_2}$, $T_1 < T_2$.

   - Acceptance probability (Jasra *et al.*, 2007):

     $$A = min\{1, \frac{\pi_i(\theta_j)}{\pi_i(\theta_i)}\frac{\pi_j(\theta_i)}{\pi_j(\theta_j)}\}. \tag{2.15}$$

In parallel tempering there are essentially two types of moves. The first is the mutation step, which simply is the within chain move we described in the previous section. This is accepted with probability given by (2.14). The other is the exchange step, a between chains move (Figure 2.2).

This Metropolis-Hastings move proposes to swap the value of two chains i and j, adjacent in terms of $t = \frac{1}{T}$, with respective temperatures $t_1 = \frac{1}{T_1}$ and $t_2 = \frac{1}{T_2}$ where $T_1 < T_2$. Suppose that the values of the two chains are θ_i and θ_j respectively, corresponding to two different sets of species. The move is accepted with probability given by equation 2.15.

When θ_j has a higher probability than θ_i, the exchange will always be accepted.

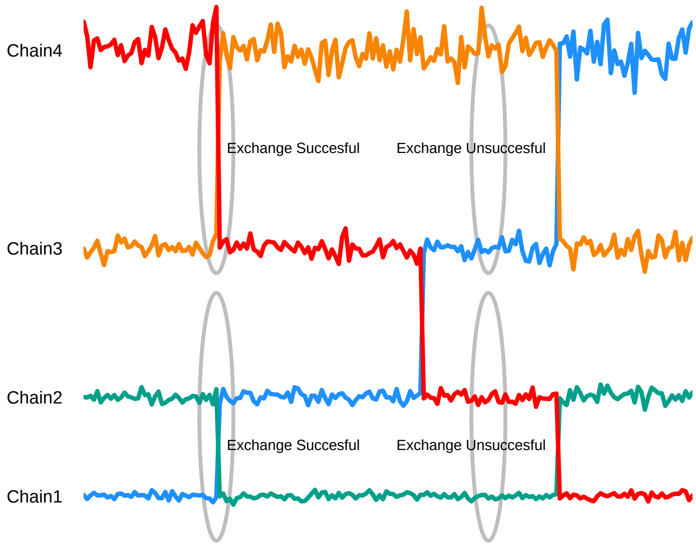


Figure 2.1: Schematic of parallel tempering. Exchanges are attempted between chains of neighboring temperatures, where Chain1 at $T_1 = 1$, $T_1 < T_2 < T_3 < T_4$.

This is simple to show considering $\pi_i(\theta_i)$ and $\pi_j(\theta_j)$:

$$\log \frac{\pi_i(\theta_j)}{\pi_i(\theta_i)} \frac{\pi_j(\theta_i)}{\pi_j(\theta_j)} = \log \frac{\pi(\theta_j)^{t_1} \pi(\theta_i)^{t_2}}{\pi(\theta_i)^{t_1} \pi(\theta_j)^{t_2}} \tag{2.16}$$

$$= (t_1 - t_2)(\log \pi(\theta_j) - \log \pi(\theta_i)). \tag{2.17}$$

Since $t_1 > t_2$ and $\log \pi(\theta_j) > \log \pi(\theta_i)$, the move is always accepted. Allowing the chains to swap states facilitates jumps between separate modes and improves the mixing rate of the cold chain, ensuring a global exploration of the model state space. Eventually hot and cold chains will progress towards a global mode. We demonstrate this by plotting the log-likelihoods for 4 tempered chains (out of 16) across iterations (Figure 2.3, based on data from clinical case 1 in chapter 5).

2.6 Bayesian model choice and model averaging

Marginal likelihood estimation has a central role in comparing different models $\{M_1, \ldots, M_m\}$, helping us to assess which model is most plausible given the data. The marginal likelihood is a weighted average of the likelihood, with the weights coming from the prior. To compute the marginal likelihood $P(X|M_k)$ for the model M_k we have to average over the parameters with respect to the prior distribution $\pi(\theta_k|M_k)$, where

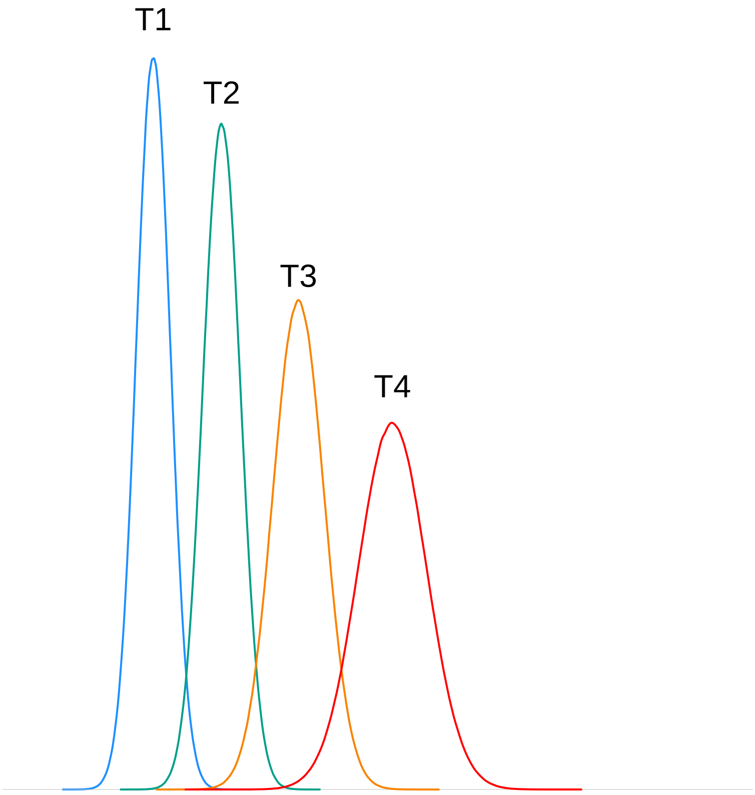


Figure 2.2: The posterior distribution at higher temperature spreads out its mass and becomes flatter. Overlap between chains at different temperatures allows for acceptance of the the exchange moves. $T_1 = 1$, $T_1 < T_2 < T_3 < T_4$.

θ_k are the model parameters:

$$P(X|M_k) = \int_{\theta_k} P(X|\theta_k, M_k)\pi(\theta_k|M_k)d\theta_k. \tag{2.18}$$

The posterior probability of the model M_k is:

$$P(M_k|X) \propto P(X|M_k)P(M_k). \tag{2.19}$$

The term $P(X|M_k)$ is the evidence for model M_k, which appears as the normalising constant in Bayes theorem 2.1. $P(M_k)$ on the other hand is the prior belief we hold for each model. It essentially expresses how plausible we thought the alternative models were before observing the data. If we were to assume that there is no strong reason to assign very different priors to the alternative models, the models would be ranked just by the marginal likelihood, otherwise by the posterior probability of the models.

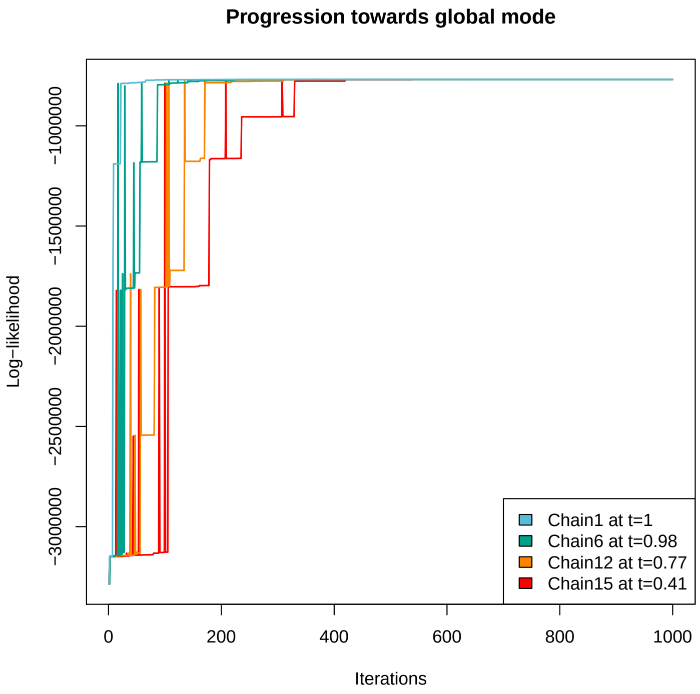


Figure 2.3: Log-likelihood traceplot of tempered distributions, where Chain1 at $T_1 = 1$, $T_1 < T_6 < T_{12} < T_{15}$.

Let us assume that we want to assess the performance of two different models M_1 and M_2. The Bayes Factor (Kass and Raftery, 1995) provides the relative weight of evidence for model M_1 compared to model M_2 and it is formulated as below:

$$BF = \frac{P(X|M_1)}{P(X|M_2)} \tag{2.20}$$

and therefore the posterior odds is simply the multiplication of the Bayes Factor with the prior odds:

$$\frac{P(M_1|X)}{P(M_2|X)} = \frac{P(X|M_1)}{P(X|M_2)} \frac{P(M_1)}{P(M_2)}. \tag{2.21}$$

Approximating the marginal likelihood is a task both difficult and time-consuming (Marin and Robert, 2008). Accounting for the uncertainty in parameters θ, a Monte Carlo approximation can be used for the marginal likelihood, by drawing independent samples from the prior to estimate $P(X)$ and averaging the likelihood. The simulation from the prior is computationally inefficient, as the majority of samples are outside the regions of high likelihood.

Importance sampling techniques can be used to reduce the variance of the estimator (Liu, 2001). If we let an MCMC sampler to explore the IS proposal distribution g, the marginal likelihood can be estimated using n sampled values $(y_1, \cdots, y_n)$ as below:

$$\hat{I} = \frac{\sum_{i=1}^{n} a_i P(X|y_i)}{\sum_i a_i} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{P(X|y_i)\pi(y_i)}{g(y_i)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(y_i)}{g(y_i)}} \tag{2.22}$$

where y_i represents the ith parameter vector sampled from the importance distribution, $P(X|y_i)$ is the likelihood computed at y_i and a_i is the importance weight for observation i computed as in equation 2.7.

The choice of the importance distribution is a crucial step that defines the stability and the accuracy of the IS estimator. In order for the method to work efficiently we want to make the estimation error as small as possible. The posterior distribution is a convenient choice as we can use the same MCMC runs we already have from the parameter estimation step. This essentially means that the marginal likelihood is estimated from MCMC samples that must be collected anyway. However using the posterior distribution or an approximation to the posterior, as the importance distribution means that the IS estimator becomes the Harmonic Mean Estimator (HME) (Newton and Raftery, 1994), as shown below:

$$\hat{I} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{P(X|y_i)\pi(y_i)}{\frac{1}{\frac{P(X|y_i)\pi(y_i)}{P(X)}}}}{\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(y_i)}{\frac{1}{\frac{P(X|y_i)\pi(y_i)}{P(X)}}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} 1}{\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(y_i)}{P(X|y_i)\pi(y_i)}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{P(X|y_i)}}. \tag{2.23}$$

The Harmonic Mean Estimator has unfortunately been one of the most popular choices for estimating the marginal likelihood, due to its simplicity. However using the HME is known to be unstable and to overestimate the marginal likelihood $P(X)$ (Xie *et al.*, 2011).

In order to overcome this issue and to make the distribution tails heavier, we perform Defensive mixture Importance Sampling (Hesterberg, 1995). This can be achieved by selecting the importance distribution g to be fairly similar to the target distribution but with heavier tails, that is “close” in shape to $\pi(y)P(X|y)$. This simple solution ensures that the weights will always be finite and in practice will not be extremely large, by using a mixture of the posterior and the prior as the IS distribution. The main idea is the incorporation of a heavy tail component in the importance function g, effectively substituting it by the mixture:

$$\lambda g + (1-\lambda)\pi, 0 < \lambda < 1 \tag{2.24}$$

where λ is close to 1, g is the approximation to the posterior using the already obtained MCMC samples and π the prior that acts as the stabilizing factor. In practice that means that the samples we use for the defensive IS estimator are generated with probability λ from g and with probability $1-\lambda$ from π. This approach is only slightly costlier in computational time compared to the typical IS. On the other hand the price we pay for the inclusion of the stabilization factor is that it increases the variance compared to using ordinary IS.

Up to this point the basis of the discussion has been model selection, however frequently there may be ambiguity over which single model to select. Using a model for further inference could imply that we are completely certain that the reported model generated the data, which is typically not the case. Bayesian model averaging (BMA) (Hoeting *et al.*, 1999) offers a solution to incorporate model uncertainty by averaging over all models. This model average is proportionally weighted by the models' posterior probabilities. The posterior probability of model M_k is:

$$P(M_k|X) = \frac{P(X|M_k)P(M_k)}{\sum_{i=1}^{m} P(X|M_i)P(M_i)}. \tag{2.25}$$

This equation (2.25) provides a way of summarizing model uncertainty after observing the data. As an example the BMA estimate of a parameter θ is (Hoeting, 2002)

$$\hat{\theta}_{BMA} = \sum_{k=1}^{m} \hat{\theta}_k P(M_k|X) \tag{2.26}$$

where $\hat{\theta}_k$ is the posterior mean for θ for model M_k.

Bayesian model averaging involves different challenges such as the computation of marginal likelihood for a large number of models or the specification of the priors for the different models. A popular approach for managing the sum in 2.25 is to explore the space of models using an MCMC.

2.7 Finite mixture model

A finite mixture model provides a flexible way to model heterogeneous data. Mixture distributions are typically used to model data where each observation has arisen from one of a number of different groups. Let us assume that the data of interest belong to one of k classes, while the individual class memberships are unavailable.

In a finite mixture model, data $X = (x_1, \dots, x_n)$ is modeled by a mixture of k fixed probability distributions (McLachlan and Peel, 2000):

$$p(x_i|\theta, k) = p(x_i|\theta) = \sum_{j=1}^{k} w_j f(x_i|\phi_j) \tag{2.27}$$

where $w = (w_1, ..., w_k)$ are the mixture weights constrained that $0 \leq w_j \leq 1$ and $\sum_j w_j = 1$. The component distributions $(f_1 = f(X|\phi_1), \dots, f_k = f(X|\phi_k))$, describe the probabilistic mechanism of generating data from each category, with $f_j(x_i) = p_{ij}$ the probability of observing x_i conditional on the assumption that it originated from category j. Let us assume that the component distributions are discrete distributions. For the remainder of the thesis when we refer to mixtures we assume these are mixtures of discrete distibutions over a finite number of categories. $\phi = (\phi_1, ..., \phi_k)$ are the component specific parameters. With θ we represent the entire set of model parameters, that is both the mixture weights and the component parameters: $(\theta_1, ..., \theta_k) = ((w_1, \phi_1), ..., (w_k, \phi_k))$.

Deriving analytically the maximum likelihood estimators or Bayes estimators is practically impossible due to the mixture model representation in 2.27. Even though conjugate priors may be used for both mixture and component parameter, the explicit representation of the corresponding posterior expectation involves the expansion of the likelihood:

$$\prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(x_i|\phi_j) \tag{2.28}$$

into k^n terms. This is computationally prohibitive for more than a few observations. Algorithms such as the Expectation-Maximization (EM) algorithm (Dempster and Laird, 1977) or the Gibbs sampler (Diebolt and Robert, 1994) address the problem of solving the likelihood equations for mixtures of distributions. We introduce both below, starting with the concept of missing data.

Before we proceed, let us assume for all following mixture model discussions that only the mixture weights are unknown (which means that the component parameters are completely specified and need not be estimated) as this will be the case when applied to the metagenomics problem of community profiling.

2.7.1 Missing data structure

In the mixture model setting each observation x_i, $1 \leq i \leq n$ is assumed to arise from a specific but unknown component of the mixture. The mixture structure is deconvoluted by the introduction of latent variables: we associate x_i with $z_i = (z_{i1}, \ldots, z_{ik})$, a k-dimensional vector indicating to which component x_i belongs.

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to class j} \\ 0 & \text{otherwise} \end{cases} \tag{2.29}$$

Therefore, each vector z_i is generated by a multinomial distribution consisting of one draw on k categories with probabilities w. We write $z_i \sim \text{Mult}_k(1; w_1, \ldots, w_k)$. The likelihood of the complete-data (the observed and missing data) $(X,Z) = (x_i, z_i, i = 1, \ldots, n)$ is:

$$p(X,Z|w) = \prod_{i=1}^{n} \prod_{j=1}^{k} (w_j p_{ij})^{z_{ij}}. \tag{2.30}$$

Integrating out the missing data $z_1, \ldots, z_n$ we get the model (2.27):

$$p(X_i|w) = \sum_{j=1}^{k} Pr(z_i = j|w) p(X_i|z_i = j, w) = \sum_{j=1}^{k} w_j p_{ij}. \tag{2.31}$$

The classification probability that observation x_i arises from $j-th$ component of the mixture, is (McLachlan and Peel, 2000):

$$\hat{z_{ij}} = Pr(z_i = j|x_i, w) = \frac{Pr(z_i = j, x_i|w)}{Pr(x_i|w)} = \frac{Pr(z_i = j|w) p(x_i|z_i = j, w)}{\sum_{j=1}^{k} Pr(z_i = j|w) p(x_i|z_i = j, w)} = \frac{w_j p_{ij}}{\sum_{j=1}^{k} w_j p_{ij}}. \tag{2.32}$$

2.7.2 Model fitting - EM

The Expectation Maximization (EM) approach to parameter estimation is a numerical optimisation procedure designed to obtain point estimates of the parameters by maximizing the likelihood (Dempster and Laird, 1977). This paper addressed the problem of solving the likelihood equations for mixtures of distributions, making it one of the first applications of EM. The algorithm is based on the missing data representation introduced in the previous section 2.7.1 and it consists of two steps. In the first step, the expected value of the missing variables z_i is computed based on $p(Z|X,w)$. In the next step we calculate the new mixing parameters w that maximize the expected complete-data log likelihood. The process is iterated until convergence.

The complete-data likelihood is given by:

$$p(X,Z|w) = \prod_{i=1}^{n}\prod_{j=1}^{k}(w_j p_{ij})^{z_{ij}} \tag{2.33}$$

and the expected complete-data log likelihood by:

$$\mathbb{E}[\ln p(X,Z|w)] = \sum_{Z\in\mathcal{Z}} p(Z|X,w)\ln p(X,Z|w) \tag{2.34}$$

where $\mathcal{Z}$ is the space of all possible values of Z, $p(Z|X,w) = \prod_{i=1}^{n} p(z_i|x_i,w)$ and $\sum_{Z\in\mathcal{Z}} p(Z|X,w) = 1$.

The EM algorithm is described below in Algorithm 2.3. It has been proved that at each iteration the likelihood is guaranteed to increase (Dempster and Laird, 1977).

2.7.3 Model Fitting - Gibbs sampling

Bayesian approaches to mixture modelling have become increasingly popular (Marin *et al.*, 2005), as they allow for probability statements to be made directly about the unknown parameters. Additionally prior beliefs can be included in the analysis. Similar to EM they also allow the complicated structure of a mixture model to be simplified through the use of latent variables. Gibbs sampler (Diebolt and Robert, 1994) is a Markov Chain Monte Carlo method particularly suited to the mixture model context.

The Gibbs sampler is based on the successive simulation of z and w. After convergence we obtain the full posterior distribution of w. A practical prior for the mixing parameters w is the Dirichlet distribution (equation 2.38), owing to its conjugate status

Algorithm 2.3 EM algorithm

- Initialization $w^{(0)}$
- At iteration t

1. **Expectation step.** Generate $z_i^{(t)}$ from $p(z_i^{(t)} = j|x_i, w_j^{(t-1)})$.

$$\hat{z_{ij}} = \frac{p(z_i = j, x_i|w)}{p(x_i|w)} = \frac{p(z_i = j|w)p(x_i|z_i = j, w)}{\sum_{j=1}^{k} p(z_i = j|w)p(x_i|z_i = j, w)} = \frac{w_j p_{ij}}{\sum_{j=1}^{k} w_j p_{ij}} \tag{2.35}$$

2. **Maximization step.** Given z_i from E-step, calculate new $w^{(t)}$ that maximize the expectation of the complete-data log-likelihood (eq.2.34).

$$w^{(t)} = \underset{w}{\operatorname{argmax}} \mathbb{E}[\ln p(X, z|w)] \tag{2.36}$$

It can be shown that this

$$w^{(t)} = \frac{\sum_{i=1}^{n} z_i^{(t)}}{N} \tag{2.37}$$

to the multinomial distribution.

$$\pi(w) = Dir(\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum\limits_{j=1}^{k} \alpha_j)}{\prod\limits_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} w_j^{\alpha_j - 1} = \frac{1}{B(\alpha)} \prod_{j=1}^{k} w_j^{\alpha_j - 1} \tag{2.38}$$

where α is positive. Generally, choosing a conjugate prior makes it possible to perform Bayesian inference in a computationally efficient manner. We can deduce that the posterior $\pi(w|z)$ (equation 2.42) is a Dirichlet with parameters $(\alpha_k + n_k)$ since $\pi(w|z) \propto \pi(z|w)\pi(w)$.

The conjugate prior $\pi(w)$ for w is the Dirichlet distribution with parameters $\alpha = \{\alpha_1 \ldots, \alpha_k\}$. Additionally the likelihood is multinomial:

$$\pi(z|w) = \frac{n!}{\prod n_k!} \prod_{j=1}^{k} w_j^{n_j}. \tag{2.39}$$

Therefore

$$\pi(w|z) \propto \pi(z|w)\pi(w) = \frac{n!}{\prod n_k!}\prod_{j=1}^{k} w_j^{n_j}\frac{1}{B(\alpha)}\prod_{j=1}^{k} w_j^{\alpha_j-1} = \frac{1}{B(\alpha+n)}\prod_{j=1}^{k} w_j^{\alpha_j+n_j-1} \tag{2.40}$$

i.e sampling from $Dir \sim (\alpha_1+n_1,\ldots,\alpha_k+n_k)$.

The algorithm is described in Alg. 2.4.

Algorithm 2.4 Gibbs Sampler algorithm

- Initialization $w^{(0)}$
- At iteration t

  1. Generate $z_i^{(t)}$ from $p(z_i^{(t)} = j|x_i, w_j^{(t-1)})$. So

     $$z_i \sim Mult(1;\hat{z_{i1}}^{(t-1)},\ldots,\hat{z_{ik}}^{(t-1)}) \tag{2.41}$$

     where $\hat{z_{ij}}$ is given by equation 2.35.

  2. Compute $n_j^{(t)} = \sum_{i=1}^{n} z_{ij}^{(t)}$.

  3. Generate $w^{(t)}$ from

     $$\pi(w|z^{(t)}) \sim D(\alpha_1+n_1^{(t)},\ldots,\alpha_k+n_k^{(t)}). \tag{2.42}$$

Chapter 3

Bayesian mixture models for metagenomic data

The methodological work presented in this chapter is my proposed approach for performing community profiling in metagenomic data. The ideas introduced in the previous chapter 2 are connected here to the community profiling problem.

metaMix considers the competing models that could accommodate our observed data, that is the BLASTx similarity between the reads and the reference proteins (Table 1.1), and compares them. The different models represent different sets of species being present in the sample. The method works on two levels of inference: in the first instance we assume a set of species to be present in the sample and we estimate this model's parameters given the data. The other level of inference is the model comparison so as to assess the more plausible model. The process is iterated in order to explore the model state space. In the following sections we describe in detail how each step of metaMix is implemented and the algorithm is summarised in Algorithm 3.1.

3.1 Model specification assuming a fixed set of species

Assuming a given set of K species from which the reads can originate, the metagenomic problem can be summarized as a mixture problem, for which the assignment of the sequencing reads to species is unknown and must be determined. The data consist of N sequencing reads $X = (x_1, \ldots, x_N)$. We have recorded the number of mismatches between pairs of translated reads and proteins as produced by BLAST (Table 3.1).

If a read matches multiple proteins from the same species, we keep only the best match. For simplicity, we assume a one to one relationship between a species proteins

Table 3.1: Sparse dissimilarity matrix - input for community profiling taken from BLAST results

	species$_1$	species$_2$	...	species$_K$
read$_1$	m$_{11}$	m$_{12}$	...	m$_{1K}$
read$_2$	m$_{21}$	m$_{22}$	...	m$_{2K}$
		...		
read$_N$	m$_{N1}$	m$_{N2}$	...	m$_{NK}$

and the species itself. Therefore, we drop the protein information and only keep a record of mismatches between each read and the different species.

Therefore for a given read x_i we can write the likelihood based on equation 2.27:

$$p(x_i|w,K) = p(x_i|w) = \sum_{j=1}^{K} w_j f_j(x_i) \tag{3.1}$$

where $w = (w_1, ..., w_K)$ represent the proportion of each of the K species in the mixture. The mixture weights are constrained such that $0 \leq w_j \leq 1$ and $\sum_j w_j = 1$. In practice for our purposes, we also add a single category (species $K+1$) which we refer to as the "unknown" category, and captures the fact that some reads cannot be assigned to any species. This may be due to the fact that the reads originate from species not included in our reference database of choice and that do not have close relatives in the database. Alternatively they may be originating from truly undiscovered species.

Additionally $f_j(x_i)$ = $P(x_i|x_i$ from species $j)$ = p_{ij} is the probability of observing the read x_i conditional on the assumption that it originated from species j. We model this probability using the number of mismatches m_{ij} between the translated read sequence i and the reference sequence j and a Poisson distribution with parameter λ for that number of mismatches:

$$p_{ij} = \frac{\text{Pois}(m_{ij};\lambda)}{l_g} \tag{3.2}$$

where l_g is the length of the reference genome, when short reads are matched to a nucleotide database. For nucleotide matching, l_g has a large impact on the probability computation. However, when matching against protein databases, the more limited heterogeneity of protein lengths results in a much smaller impact of the length parameter. In addition, incomplete annotation can potentially make the inclusion of protein length problematic for the p_{ij} computation. Consequently, for protein matched sequences, we simply defined our p_{ij} as:

$$p_{ij} = \text{Pois}(m_{ij};\lambda) \tag{3.3}$$

Therefore for a given set of K species, the p_{ij} probabilities are completely specified (Table 3.2) and only the mixture weights need to be estimated.

Table 3.2: Sparse matrix of p_{ij} probabilities

	species$_1$	species$_2$	...	species$_K$
read$_1$	p$_{11}$	p$_{12}$	...	p$_{1K}$
read$_2$	p$_{21}$	p$_{22}$	...	p$_{2K}$
		...		
read$_N$	p$_{N1}$	p$_{N2}$	...	p$_{NK}$

Combining the above we conclude that when we know the set of species, the mixture distribution gives the probability of observing read x_i: $\sum_{j=1}^{K} w_j p_{ij}$ namely equation (2.27). We therefore write the likelihood of the dataset X as a sum of K^n terms:

$$P(X|w) = \prod_{i=1}^{n}[\sum_{j=1}^{K} w_j p_{ij}]. \tag{3.4}$$

3.1.1 Choice of parameter values

We have currently set λ to 0.03, that is we would expect by chance three mismatched nucleotides (respectively three mismatched amino acids) per 100 nucleotides (amino acids), including both sequencing errors and mutations. The λ parameter is currently non tunable but this will be amended in the next metaMix release. However the users can control how divergent species are treated by metaMix by increasing or decreasing the value of $p_{i\text{unknown}}$. This is the default probability for reads to be generated by the unknown category, which is the collective of unknown/undiscovered taxa.

In order to choose a value for the $p_{i\text{unknown}}$ in the protein comparison context, we tested different values and settled on 10^{-6}. For $\lambda = 0.03$ the reads that have fifteen or more mismatches per 100 amino acids (∼85% similarity) to the proteins in the reference database, will have p_{ij} smaller than $p_{i\text{unknown}} = 10^{-6}$. Even though the contribution of each species j to the likelihood is defined by both the p_{ij} and its relative mixture weight w_j as seen by equation 3.4, in the most typical scenarios species that are supported only by low similarity reads will not be retained in the species profile and therefore these reads will be assigned to the unknown category. We found this value to offer a good balance, allowing us to consider divergent matches but at the same time to retain a level of stringency in the results. However, users can set different values for the unknown p_{ij} increasing or decreasing the effect non-exact matches have in the results.

Finally, if the users wish to investigate the nature of their unclassified reads, they can retrieve these from the results for follow up analyses.

3.2 Estimation of mixture weights

Assuming we know the set of species present, we wish to estimate species abundances, that is the mixture weights. The probabilistic assignment of reads to species involves the expansion of the likelihood into K^n terms which is computationally infeasible through direct computation. An efficient estimation can be performed by the introduction of unobserved latent variables that code for the read assignments. As discussed previously, either the Expectation-Maximization (EM) algorithm or the Gibbs sampler could be used to estimate the mixture weights w. EM returns a point estimate for w while the Gibbs sampler the distribution of w. Both methods were implemented and provided comparable results, but we chose the EM for computation speed.

3.3 Marginal likelihood estimation

Each combination of species, or else each different community profile, corresponds to a finite mixture model for which the marginal likelihood can be estimated. Let us assume we wish to compare the different models $\{M_1, \ldots, M_m\}$. The marginal likelihood $P(X|M_k)$ for the mixture model M_k with parameters θ_k is given by equation (2.18). The posterior probability of the model is given by $P(M_k|X) \propto P(X|M_k)P(M_k)$.

The prior $P(M_k)$ can be specified depending on the context and the basis of our interpretation is that parsimonious models with a limited number of species are more likely. Thus, in this Bayesian framework, our default prior uses a penalty limiting the number of species in the model: $P(M_k) \propto penalty^{(\text{number of species in } M_k)}$. We approximate this penalty factor based on a user-defined parameter `r` that represents the species read support required by the user to believe in the presence of this species.

We compute the logarithmic penalty value as the log-likelihood difference between two models: model M_{unknown} which is our starting point when we have no knowledge about which species are present and therefore all N reads come from the "unknown" category ($p_{ij} = 10^{-6}$) and model $M_{\mathtt{r}}$ where r reads have a perfect match to a species ($p_{ij} = 1$) and the remaining $N - \mathtt{r}$ reads belong to the "unknown" category:

$$\log penalty = \log P(M_{\text{unknown}}|X) - \log P(M_{\mathtt{r}}|X). \tag{3.5}$$

For DNA sequence analysis, the p_{ij} probabilities for the r reads originating from this unspecified species are approximated by 1/(median genome length in the reference database). This read support parameter reflects the number of unique reads required to support the hypothesis that a species is present.

From now on, when we refer to the marginal likelihood, we mean the marginal likelihood for a specific model and we forego conditioning on the model M_k in the notation. Additionally, in our mixture model p_{ij} are completely specified, therefore the model parameters θ_k are solely the mixture weights w. Hence the marginal likelihood equation (2.18) described in the previous chapter becomes:

$$P(X) = \int_w P(X|w)\pi(w)dw \overset{(3.4)}{=} \int_w \prod_i [\sum_j w_j p_{ij}]\pi(w)dw. \tag{3.6}$$

We implemented the Defensive mixture IS procedure for the estimation of the marginal likelihood, by using a mixture of posterior and prior as the IS distribution. We chose the Defensive mixture IS technique for the relatively simple implementation compared to other approaches. This is crucial as we perform this approximation numerous times, for every species combination we consider.

We first approximate the posterior distribution $P(w|X)$ with a normal multivariate distribution g. We use the output of the Gibbs sampler to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ for the parameters w and setting these as the parameters for the multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We then incorporate a heavy tail component in the importance function g by using the prior distribution $\pi(w)$. Finally, we generate n samples $1 \leq i \leq n$, for the defensive IS estimator which are generated with probability $\lambda = 0.95$ from g and with probability $1 - \lambda = 0.05$ from π.

However the goal of this work is to deliver results within an actionable time-frame in a clinical setting. We wish to speed up the computation without compromising the accuracy and the sensitivity of the results. For that reason, we use a point estimate of the marginal likelihood by means of the Expectation-Maximization (EM) algorithm. The different approaches were used on the benchmark dataset. The results were compared and are discussed in chapter 4, section 4.3.6. The resulting taxonomic assignment as well as the species relative abundance estimates were similar between them, with the EM approach resulting in a 13-fold speed increase.

3.4 Model comparison: exploring the set of present species

We use a Monte Carlo Markov Chain (MCMC) to explore the set of present species of size $2^S - 1$, where S is the total number of potential species. In practice we observe that S can be greater than 1,000. The MCMC must explore the state-space in a clinically useful timespan. Therefore we reduce the size of the state-space, by decreasing the number of S species to the low hundreds. We achieve this by fitting a mixture model with S categories, considering all potential species simultaneously. Post fitting, we retain only the species categories that are not empty, that is categories that have at least one read assigned to them.

Let us assume that at step t, we deal with a set of species that corresponds to the mixture model M_k. At the next step $(t+1)$, we either add or remove a species and the new set corresponds to the mixture model M_l. The step proposing the model M_l is accepted with probability:

$$A(M_k \rightarrow M_l) = min\{1, \frac{P(X|M_l)^{(t+1)}P(M_l)}{P(X|M_k)^{(t)}P(M_k)} \frac{q(M_l \rightarrow M_k)}{q(M_k \rightarrow M_l)}\} \tag{3.7}$$

where $q(M_l \rightarrow M_k)$ is the probability of proposing model M_k when currently at model M_l. In other words, this is the probability of adding or removing the species to the M_k set of species that took us to the M_l set of species. If the step is accepted, then the chain moves to the new proposed state M_l. Otherwise if not accepted, the chain's current state becomes the previous state of the chain, which means that the set of species remains unchanged.

metaMix outputs log-likelihood traceplots so that the user can visually inspect the mixing and the convergence of the chain. The original version of metaMix was based on the use of a single chain MCMC. The likelihood traceplots we would obtain from metaMix on different datasets were indicative of poor mixing of the chain (Figure 3.1a). As discussed in the previous chapter this occurs when different modes are separated by regions of low posterior density. This suggested that simple MCMC does not efficiently explore the complex model state space we typically work with in the metagenomic field and that such a chain would take a long time to reach equilibrium.

In the parallel tempering MCMC setting, each chain simulates from the posterior distribution $P(M_k|X)$=$g(M_k)$ raised to a temperature $\{t_1 = \frac{1}{T_1}, t_2 = \frac{1}{T_2}, \dots, t_{max} = \frac{1}{T_{max}}\}$ where model M_k comes from a collection of models $\{M_1, \dots, M_m\}$ each corresponding to a different set of species. The mutation move is defined as above by 3.7 and the exchange move between two neighboring chains k_1 and k_2 with values M_{k_1} and M_{k_2} respectively, corresponding to two different sets of species, becomes:

$$A = min\{1, \frac{g_{k_1}(M_{k_2})}{g_{k_1}(M_{k_1})} \frac{g_{k_2}(M_{k_1})}{g_{k_2}(M_{k_2})}\}. \tag{3.8}$$

The chains are adjacent in terms of $t = \frac{1}{T}$, with respective temperatures $t_1 = \frac{1}{T_1}$ and $t_2 = \frac{1}{T_2}$ where $T_1 < T_2$.

Using the metaMix output for one of the clinical datasets, we plotted the log-likelihoods for the second half of 3000 iterations for a single chain MCMC (Figure 3.1a). We compare this with the parallel tempering MCMC output, plotting for the cold chain the second half of 1000 iterations (Figure 3.1b). It is easy to observe that the cold chain from the PT MCMC exhibits better mixing, as it moves around the state space with ease. Aditionally a snapshot of the traceplots during iterations 500-1000 reveals that the PT MCMC is moving between areas of greater likelihood compared to the single chain MCMC (Figure 3.2).

We have found that discarding the first 20% of the iterations as burn-in is enough and we have therefore set this as the default setting. We concentrate on the remaining 80% to study the posterior distribution over the model choices. We want to incorporate model uncertainty and thus we perform Bayesian model averaging as described before.

A useful summary is the posterior probability that a specific species is present in the community profile post burn-in. The different species combinations may be represented by a vector of binary variables, $(S_1, \dots, S_k)$ where S_j is an indicator for the inclusion of species j under a specific model. We write this posterior probability as

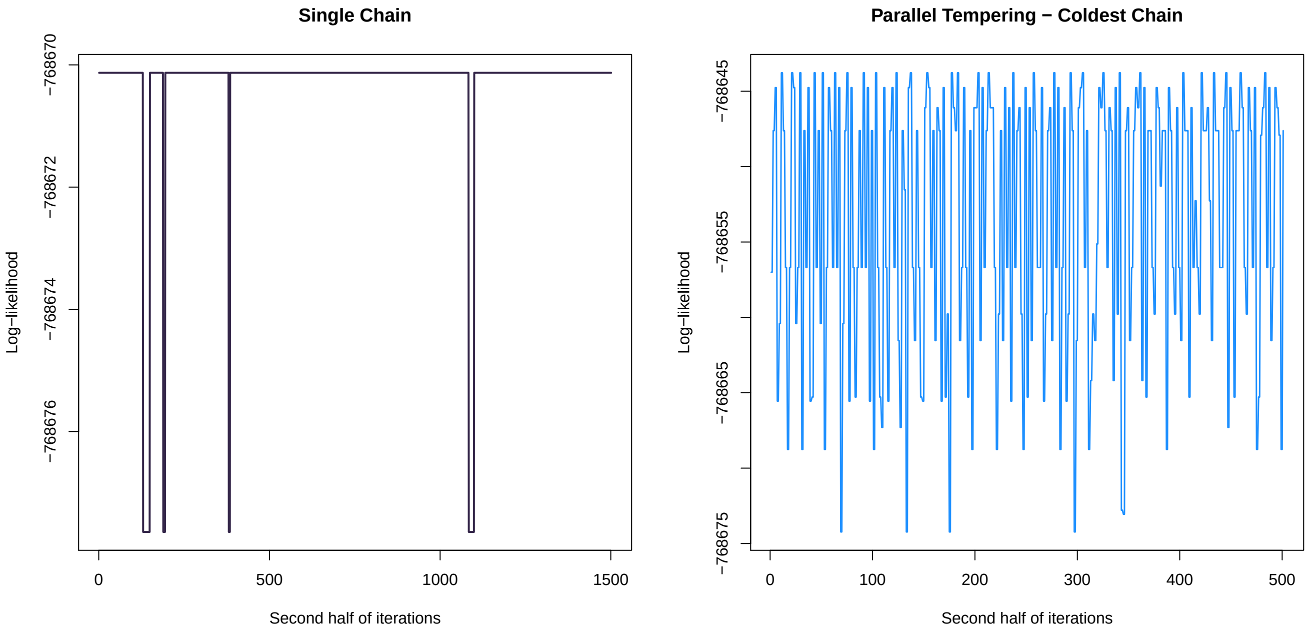


Figure 3.1: **a.** Log-likelihood trace plot for single chain MCMC and **b.** for PT chain at temperature T=1.

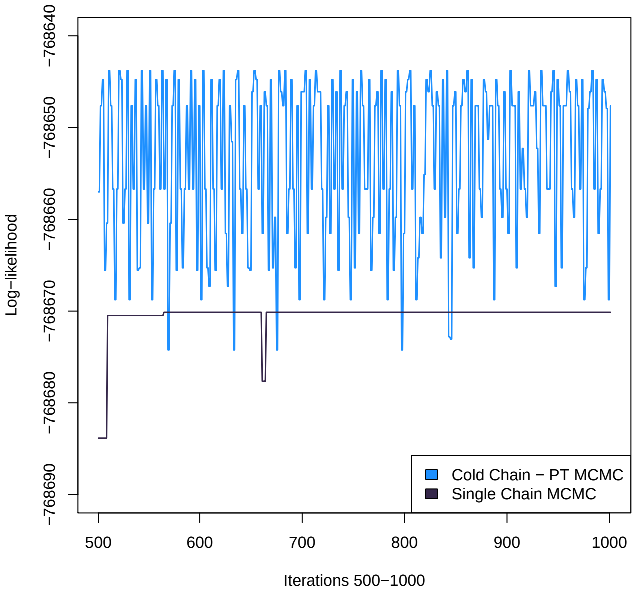


Figure 3.2: Traceplot snapshot at 500-1000 iterations for single chain MCMC and cold chain PT MCMC

$P(S_j = 1|X)$ and it can be obtained by summing the posterior model probabilities over all models where species j is present.

$$P(S_j = 1|X) = \sum_{q=1}^{m} p(M_q|X)p(S_j = 1|M_q,X). \tag{3.9}$$

The species presence information is recorded by metaMix as in the example in Table 3.3.

	S_1	S_2	$\cdots$	S_j	$\cdots$	S_k	
model M_1	0	1	$\cdots$	0	$\cdots$	0	$p(M_1|X)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
model M_q	1	1	$\cdots$	0	$\cdots$	0	$p(M_q|X)$
model $M_{(q+1)}$	1	0	$\cdots$	0	$\cdots$	0	$p(M_{q+1}|X)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
model M_n	1	1	$\cdots$	1	$\cdots$	0	$p(M_n|X)$

Table 3.3: metaMix recorded information on species presence at each MCMC iteration.

We can thus summarize appropriately the posterior distribution and answer the important questions of interest. Examples of such questions include: what species have probability p or greater being included in the set of present species? what is the probability of having the n specific closely related strains in the set of present species? Depending on the biological context, one may ask numerous similar or other case-specific questions. Finally, metaMix also outputs Bayes Factors to quantify the evidence in favour of each species:

$$\log_{10} BF = \log_{10} \frac{P(X|M_{\text{species present}})}{P(X|M_{\text{species absent}})}. \tag{3.10}$$

metaMix consists of four steps which are summarised below in Algorithm 3.1 and will be demonstrated in the following chapter 4 using a toy example.

3.5 Practical considerations

3.5.1 Number of chains and tempering scheme

The optimal number of chains used in the parallel tempering is not obvious. The main idea is that the number of chains used must be large enough to ensure successful swaps between all neighboring chains. The first limitation is the number of chains we can run

Algorithm 3.1 metaMix algorithm

Step 1. Compute p_{ij} generative probabilities using BLAST similarity, using equation 3.2 or 3.3.

Step 2. Fit mixture model with all potential species as categories, typically $k > 1000$.
Keep for subsequent MCMC exploration the non-empty categories, i.e species that have at least one read assigned to them.

Step 3. MCMC exploration of species space using multiple parallel tempered chains.

- Initialisation of chain:
  At iteration t=0, assume all reads come from the unknown category. At iteration t=1, add a species.
- At iteration t:
  - Let us assume we deal with a set of species that corresponds to the mixture model M_k. At the next step $(t+1)$, we either add or remove a species and the new set corresponds to the mixture model M_l. The step proposing the model M_l is accepted with probability 3.7:

    $$A(M_k \rightarrow M_l) = min\{1, \frac{P(X|M_l)^{(t+1)}P(M_l)}{P(X|M_k)^{(t)}P(M_k)} \frac{q(M_l \rightarrow M_k)}{q(M_k \rightarrow M_l)}\}$$

  - After each mutation step, attempt exchange move between neighboring chains k_1 and k_2. Accept with probability 3.8:

    $$A = min\{1, \exp\{(t_1 - t_2)(\log \pi(M_{k_1}) - log\pi(M_{k_2}))\}\}$$

    where t_1 is the temperature for chain k_1 and t_2 the temperature for chain k_2.

Step 4. Compute the posterior probabilities of species being present by model averaging.

on a computer. The computing facility we use for all analyses is the UCL Computer Science cluster. The availability of suitable machines as well as considerations towards minimizing the queuing time of the submitted jobs, results in choosing $N = 12$ chains.

The choice of temperature values is motivated by the fact that these must not be too far apart, so that exchange of values between the chains can occur. Additionally the maximum value must be high enough so that no chains become trapped in local minima, hence allowing for global moves. We implemented a power decay heating scheme:

$$t_n = (t_{n-1} - K)^{\alpha}, \text{ where } n = 2, \ldots, N, \, K \in (0,1), \, \alpha > 1 \text{ and } t_1 = 1 \tag{3.11}$$

and using $K = 0.001$ and $\alpha = 3/2$ we achieve a slowly heating sequence of chains with a lot of chains similar to the target.

We find that for $N < 10$ the maximum temperature is not very high, hindering a quick global exploration. Ideally we would prefer to run 14 to 20 chains but given our computing constraint $N = 12$ performs satisfactorily.

3.5.2 Number of iterations

Given the described setup of our Parallel Tempering, we find that for metaMix to produce reasonable results there is a minimum requirement for $(5 \times \text{number of potential species})$ MCMC iterations for each chain. A greater number of iterations would naturally help achieve smaller errors. The user can visually inspect the log-likelihood traceplots to assess mixing and convergence of the cold chain and may also increase the iterations number if necessary.

Chapter 4

The metaMix R package

This chapter provides an overview of the features implemented in the metaMix R package. First, the significant functions are demonstrated with a toy example. The performance of metaMix is then assessed using metagenomic benchmark datasets where the ground truth is known. Finally metaMix is compared to other community profiling methods discussed in chapter 1.

4.1 An overview of the metaMix R package

metaMix is the R package implementing the ideas introduced in the previous chapter. It uses finite mixture models coupled with PT MCMC in order to identify the set of species most likely to be present in a metagenomic community. metaMix also estimates their relative abundances. Different competing models that could accommodate the observed data, that is the BLASTx similarity between the read sequences and the reference protein sequence, are considered and compared. The final output is the probabilistic community profile as well as supporting plots, such as log-likelihood traceplots as well as cumulative histogram plots of the classification probabilities for each species in the summary profile. Even though the implementation of the ideas is computationally intensive and requires a supercomputer, the following guide to metaMix uses a toy example where all the steps can be performed on a single machine.

4.1.1 Installation

metaMix has the following package dependancies: `Rmpi`, `data.table`, `Matrix`, `gtools` and `ggplot2`. `Rmpi` provides the interface to openMPI (Message Passage Interface). The user can check whether openMPI is installed on their computer using the command `mpirun`. More information can be found here http://www.open-mpi.org/software/ompi.

4.2 Demonstration of functions with toy example

The starting point is to obtain the sequence similarity between a query and a target sequence. This can be done with the homology tool BLAST. Both nucleotide and amino acid comparisons are supported. The metaMix demonstration below uses amino acid similarities, i.e the input file is generated by BLASTx.

4.2.1 Step1

During the first step, metaMix estimates the generative probabilities p_{ij} based on the amino acid similarity between the translated read sequence and the proteins in the reference database.

Default BLAST output

The default output tabular file is supported, obtained using `-outfmt 6` in the BLAST command: `blastx -db referenceDB -query input.fa -outfmt 6 -max_target_seqs 10`. The default output file has the following fields: `Query ID`, `Subject ID`, `Identity`, `Alignment Length`, `Mismatches`, `Gap Openings`, `Query Start`, `Query End`, `Subject Start`, `Subject End`, `E-value`, `Bit Score`. metaMix needs information on the read lengths as well as a file mapping the gi identifiers to the taxon identifiers. These are not included in the default output of BLAST and need to be provided as additional arguments.

```
>library(metaMix)
###Location of input files.
>datapath <- system.file("extdata", package="metaMix")
>blastOut.default<-file.path(datapath, "blastOut_default.tab")
>read.lengths<-file.path(datapath, "read_lengths.tab")
>read.weights<-file.path(datapath, "read_weights.tab")
taxon.file<-file.path(datapath, "gi_taxid_prot_example.dmp")
>read.table(read.lengths, nrows=2, sep="\t")
## read   read.length
NODE_427    209
```

```
NODE_428    162
>read.table(read.weights, nrows=2, sep="\t")
##read   read.weight
NODE_476    10
NODE_524    26
>read.table(taxon.file, nrows=2, sep="\t")
## GI_id   taxon_id
9625360    10849
9625363    10849
```

Custom BLAST output

Alternatively, metaMix accepts a custom BLAST output file that has already incorporated the read lengths and the taxon identifiers. This custom file has the following fields: Query ID, Query Length, Subject ID, Subject Length, Mismatches, Bit Score, Alignment Length, %Identity, E-value, Taxon ID. It is produced by the following BLAST command: blastx -db referenceDB -query input.fa -max_target_seqs 10 -outfmt "6 qacc qlen sacc slen mismatch bitscore length pident evalue staxids".

The p_{ij} probabilities are estimated by the generative.prob() function:

```
  blastOut.custom<-file.path(datapath, "blastOut_custom.tab")
  step1 <-generative.prob(blast.output.file = blastOut.custom,
  contig.weight.file=read.weights,
  blast.default=FALSE,
  outDir=NULL)
```

where blast.default denotes usage of the BLAST default output (TRUE) or the custom output specified above (FALSE). The value for the blast.output.file argument is the tabular BLASTx output file. The argument contig.weight.file can be omitted when working with unassembled reads, as the weight is set by default to be 1 - same for all reads. However if an assembly step has been performed as in this example, information on the number of reads that make up each contig needs to be provided. This will be a two column tab-separated file, where the first column is the contig identifier and the second the number of reads. Finally outDir is the directory where the results are written and where an object from each step is saved. When it is set to NULL no objects will be saved.

On the other hand, when working with the default BLAST output the command would be the following:

```
step1 <-generative.prob(blast.output.file = blastOut.default,
                        read.length.file=read.lengths,
                        contig.weight.file=read.weights,
                        gi.taxon.file = taxon.file,
                        blast.default=TRUE,
                        outDir=NULL)
```

The information missing from the BLAST file is now provided with two extra arguments: `read.length.file` can either be the file mapping each read to its sequence length or a numerical value, representing the average read length (default value=100). `gi_taxid_prot.dmp` is a taxonomy file, mapping each protein gi identifier to the corresponding taxon identifier. It can be downloaded from ftp://ftp.ncbi.nih.gov/pub/taxonomy/gi_taxid_prot.dmp.gz.

The function `generative.prob` creates a list of five elements. The first element is a sparse matrix `pij.sparse.mat` where each row corresponds to one read and each column to a species. The value of the cell is the generative probability p_{ij}. The second element is `ordered.species` containing all the species that correspond to the proteins in the BLASTx output file. Finally the `read.weights`, `gen.prob.unknown` and `outDir` are the other three elements of the list `step1`, carried forward to be used in the second step.

```
###The resulting list consists of five elements
names(step1)
[1] "ordered.species" "pij.sparse.mat" "read.weights" "outDir"
[5] "gen.prob.unknown"
### There are that many potential species in the sample:
nrow(step1$ordered.species)
[1] 224
### The sparse matrix of generative probs.
step1$pij.sparse.mat[1:2,c("374840", "258", "unknown")]
```

```
5 x 3 sparse Matrix of class "dgCMatrix"
                              374840    258    unknown
@M01520:37:13805:1480_1:N:0:1 7.366e-01  .     1e-06
@M01520:37:16186:1480_2:N:0:1 9.389e-01  .     1e-06
```

4.2.2 Step2

Theoretically the next step would be the state space exploration with the PT MCMC. In practice, the number of all potential species S is large and this step works on reducing the size of the species pool from the thousands to the low hundreds.

In this simple example there are only 224 organisms and thus step2 fits a mixture model with 224 categories, considering all the potential species simultaneously. Post fitting only the non-empty categories are retained for the MCMC exploration, that is the species that have at least one read assigned to them. The required argument for the `reduce.space` function is simply the list created in the first step using the `generative.prob` function.

```
>step2 <- reduce.space(step1=step1)
##These are the elements of the step2 list.
>names(step2)
[1] "outputEM" "pij.sparse.mat" "ordered.species" "read.weights"
[5] "outDir" "gen.prob.unknown"
## After this approximating step, there are 7 potential species in the sample:
>nrow(step2$ordered.species)
[1] 7
## And these are:
>step2$ordered.species
taxonID  countReads  samplingWeight
374840      1888       0.165017
28090       93         0.034611
13690       62         0.023074
645687      46         0.017119
```

The reduced species pool consists solely of 7 potential taxa down from 224. In the typical scenario the species reduction is from the thousands to the hundreds.

4.2.3 Step3

In this step, the different models are considered and compared. The space exploration by the parallel tempering MCMC is implemented by the function `parallel.temper`. The required argument is the list created in the second step using the `reduce.space` function. An important optional argument of this function is `readSupport`. As mentioned in the previous chapter, the default model prior uses a penalty limiting the number of species in the model. The penalty factor is approximated based on `readSupport`, which represents the species read support required from the user in order to believe in the presence of a species in the sample. The default value is 10.

```
>step3<-parallel.temper(step2=step2)
##These are the elements of the step3 list.
>names(step3)
## [1] "result" "duration"
## Steps MCMC took during some iterations.
>step3$result$slave1$record[10:15,]
Iter Move Candidate Species 374840 2 28090 13690 645687 258 1035 unknown
    logL
10 Remove 645687 1 0 1 1 1 0 0 1 -2988
11 Remove 374840 1 0 1 1 1 0 0 1 -2988
12 Add 2 1 1 1 1 1 0 0 1 -3070
13 Add 258 1 1 1 1 1 0 0 1 -3070
14 Remove 13690 1 0 1 0 1 0 0 1 -3505
15 Add 2 1 0 1 0 1 0 0 1 -3505
```

For each parallel chain, the MCMC trajectory has been recorded. The record includes information on what species were proposed and which were accepted or rejected throughout the iterations. For example at iteration 10, removing species 645687 was proposed but not accepted, as denoted by the 1 in the column 645687. Between iterations 13 and 14 an exchange of the sets of species between Chain 1 and Chain 2 occurred. At iteration 13 species 2 was present, while at the next one, it is no longer there. That means that the attempt at swapping the values of the two neighboring chains was successful. This information is also recorded, i.e how many swaps were attempted and how many accepted.

4.2.4 Step4

Having explored the different possible models, the final step is to compute the posterior probabilities for each species by performing model averaging. The MCMC model choices for Chain 1 are used to produce a probabilistic summary for the presence of the species.

The required arguments are the lists created in the second and third steps, using the `reduce.space` and the `parallel.temper` functions. Finally, the taxonomy names file 'names.dmp' which can be downloaded and extracted from ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz has to be provided.

```
## Location of the taxonomy names file.
taxon.file<-file.path(datapath, "names_example.dmp")

step4<-bayes.model.aver(step2=step2,
                        step3=step3,
                        taxon.name.map=taxon.file)
##These are the elements of the step4 list.
>names(step4)
[1] "result" "pij.sparse.mat" "presentSpecies.allInfo"
[4] "output100" "assignedReads" "classProb"
##This is the species summary
>print(step4$presentSpecies.allInfo)
taxonID        scientName                 finalAssignments   poster.prob
374840     Enterobacteria phage phiX174       2419               1.0
28090      Acinetobacter lwoffii              93                 0.9
unknown    unknown                            66                 1.0
13690      Sphingobium yanoikuyae             62                 0.9
645687     Astrovirus VA1                     46                 1.0
```

In the resulting profile there are four species and the unknown category. In this step supporting plots that help the user further assess the results are generated. These include log-likelihood traceplots as well as cumulative histogram plots of the classification probabilities for each species in the summary profile.

4.3 Benchmark datasets

The performance of metaMix on datasets where the ground truth is known was examined to compare the metaMix estimates with the real values. metaMix was applied on a popular benchmark dataset where the exact community composition and read assignment are specified. metaMix results are compared with the ones produced by two other similarity-based community profiling methods, MEGAN version 5.3 and Pathoscope 2.0. The similarity-based aspect and more specifically their flexibility to work with BLASTx output makes them better candidates for viral discovery that is the focus in this thesis, compared to composition-based methods. From the mixture model methods, we have chosen Pathoscope. Default parameters were used for all methods, unless stated otherwise.

The FAMeS artificial datasets (`http://fames.jgi-psf.org/description.html`) are mock metagenomic community datasets (Mavromatis *et al.*, 2007), composed of randomly selected real Sanger shotgun sequecing reads (average length≈800bp) from the original sequencing projects of 113 microbial genomes (Integrated Microbial Genomes database (Markowitz *et al.*, 2012)). They are a popular choice to use as benchmark datasets for various metagenomics methods. Their suitability stems from the fact that the number of species that form the metagenomic community is known as well as their relative abundances. The FAMeS datasets have been designed to model real metagenomic communities in terms of complexity and phylogenetic composition.

For the metaMix output, we reported organisms with a posterior probability greater than 0.8 (default). The metaMix read support parameter `r`, which essentially sets the sensitivity/specificity of the method, has an impact on the number of reported species. A large `r` value can result in the method merging together strains that are differentiated by fewer reads than `r`. On the other hand a low `r` can have the opposite effect, whereby the methods splits a strain into two or more strains, by moving a few reads from one strain to a very similar one with which they have equally good matches.

The user's choice for this key parameter `r` should be informed by the biological context. As an example, for the typical human clinical sample where the sample collection might have occurred some time after the infection has taken place, a low value in order to adopt a sensitive approach is reasonable. Hence, for viral identification in human clinical samples, a low and sensitive value ($\mathtt{r} = 10$) is a reasonable choice. In a highly complex environmental metagenomic community where there is a plethora of species of similar abundances, the choice becomes less straightforward especially in the case of closely related strains. We set the default value for general community profiling in environmental samples to be $\mathtt{r} = 30$. We also compare the output of metaMix for different values of this parameter as well as for different posterior probability cutoffs.

4.3.1 metaMix comparison to MEGAN, Pathoscope

There are three FAMeS datasets: simHC, simMC, simLC corresponding to high, medium and low complexity of the metagenomic community respectively. Complexity in this instance means that the communities differ in their relative abundances setup (Figure 4.1). The low complexity community (simLC) consists of reads from mostly one dominant population (28%) while the remaining 112 taxa have significantly fewer reads. In the medium complexity community (simMC) there is more than one dominant population (18%, 13%, 9%) and also similarly to simLC, the remaining taxa are of lower abundance. Finally, the high complexity dataset (simHC) lacks dominant populations and all taxa are of low abundance (below 3%). Most of the taxa are different species while there are instances of different strains within the same species group (∼10 instances of strains within *Burkholderia cenocepacia*, *Rhodopseudomonas palustris*, *Xylella fastidiosa*).

We first discuss in detail the results of the three methods for simHC, the highest complexity dataset. simHC consists of 113 bacterial taxa, most of them distinct species (and some instances of strains from the same species) with similar abundances and no dominant population. The lowest abundance is 255 reads out of 118,000 reads. We then summarise the results for the other two mock communities, simLC and simMC. The bioinformatics processing in this instance consisted of a BLASTn comparison to all NCBI bacterial genomes (`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz`). The number of genomes mapped, retrieved from the the BLASTn output was ∼2,500. As discussed below, metaMix outperforms Pathoscope and MEGAN in the community profiling task and consequently in the relative abundance estimation (Table 4.2).

metaMix

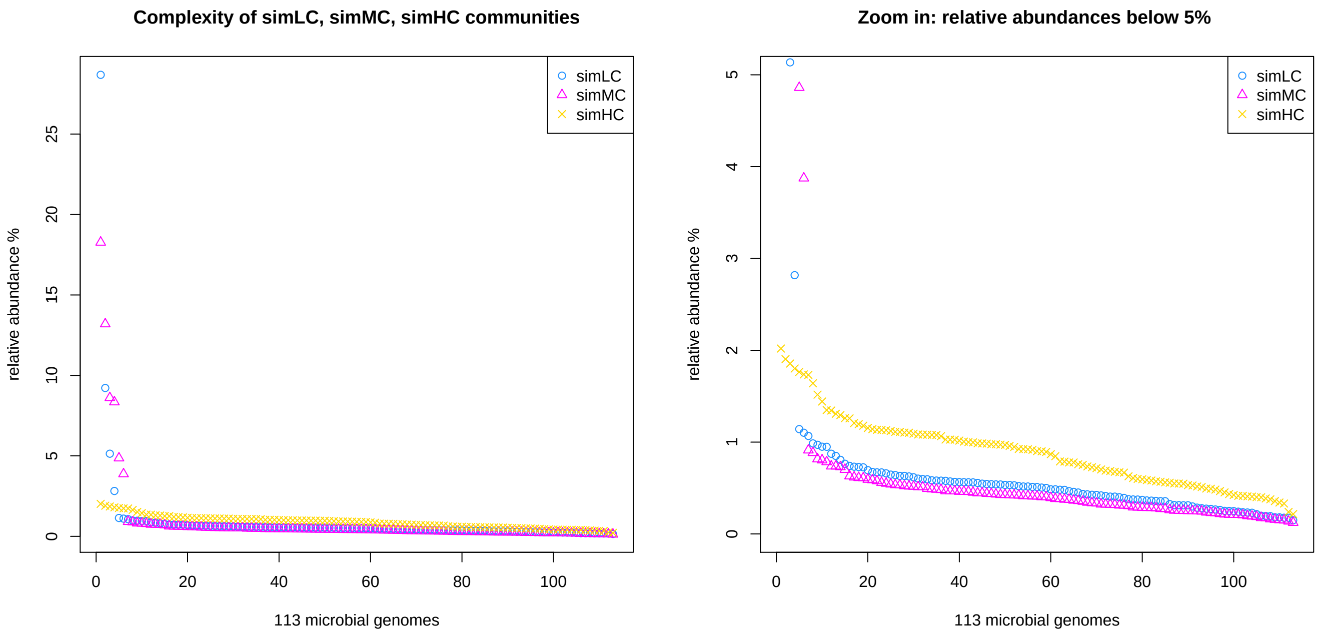


Figure 4.1: Complexity of FAMeS simLC, simMC and simHC mock communities.

To limit the complexity of the fit, the two step procedure described previously was used. First, we fitted the mixture model with the complete set of 2,500 species using EM with a limited run length of 500 iterations. Based on this analysis, we identified 1,312 species supported by at least one read and explored this state space. To limit the computational time, we also considered a stronger approximation, including only the 374 potential species supported by at least 10 sequencing reads. Both approaches generated similar results, albeit the more complex one with 1,312 potential species required the quadruple of the computation time (12 hours for 6,560 iterations instead of 3 hours for 1,870 iterations). metaMix identified 116 species, detecting successfully all the members of the metagenomic community. These were detected on the strain level except in four instances where a different strain of the same species, or different species within the same genus was detected. Four species were identified and not in the simulated dataset, hence can be considered as false positives (Table B.2).

In order to assess the variability of metaMix results, we ran the analysis 25 times changing the random seed. We report the number of species detected, the sensitivity and specificity as well as relative abundance estimate measure errors, at various posterior probability cutoffs (Table 4.1). We summarise the resulting community profile based on one of these runs in Table B.1.

Table 4.1: simHC community: number of species detected by metaMix as well as sensitivity, specificity, AVGRE, RRMSE for metaMix at various posterior probability cutoffs (default in bold font) . The results are average values based on 25 runs.

Cutoff	0.9	0.8	0.7	0.6	0.5
Sensitivity (mean)	99.82	**99.96**	99.96	100	100
Sensitivity (sd)	0.0036	**0.0017**	0.0017	0	0
Specificity (mean)	99.86	**99.82**	99.77	99.73	99.70
Specificity (sd)	0.0004	**0.0004**	0.0005	0.0003	0.0001
RRMSE	16.69	**16.85**	16.73	17.50	17.48
AVGRE	8.20	**8.31**	8.16	8.60	8.56
# Species -median	115	**116**	117	118	119
# Species - s.d	1.2	**0.9**	1.2	0.7	0.3

Pathoscope

Pathoscope identified 47 species. Of these 45 are members of the metagenomic community. 42 are the exact same strain, while 3 are either the same species but different strain or same genus but different species. However it fails to detect 68 species that are actually present in the mixture. Tuning the parameter that enforces the parsimonious results (any thetaPrior greater than 10), thereby removing the unique read penalty, Pathoscope behaves as a standard mixture model and identifies 165 species (Table 4.2). With these settings, it identifies all but one members of the community (Table B.1). The organisms are identified at the strain level, except in three instances where it identified different species within the same genus. The major interpretation issue is the presence of a long tail of species (54 species) that are actually not present in the mixture (Table B.2). Pathoscope produced the results in one minute.

MEGAN

MEGAN identified 232 taxa. It discovered all original species of the community on the strain level, except for 9 instances where it identified the lowest common ancestor (LCA). Aside from the lack of strain or species specificity for 8% of the community members, the main issue is the long tail of false positives appearing in the results, that is MEGAN exhibits low specificity (Table 4.2). In the species summary provided by MEGAN, there are 119 taxa (species or higher order) which are not actually present, but supported by a sufficient number of reads (default value: 50 reads) for MEGAN to include these in the output. It finished the computations in less than one minute. To lower the false positive rate, we also filtered the BLAST results prior to MEGAN analysis, imposing stringent E-value and similarity cutoffs, to assess how these affect the MEGAN performance. An E-value $< 1\text{E-}10$ removed only 9 entries from the results, requiring similarity greater than 90% removed only 5, while both filters resulted in 208 taxa in the summary results.

4.3.2 Relative abundances

The primary aim for metaMix is to be a diagnostic tool and to answer whether a species is present or absent from the mixture we study. As a secondary aim, we are also interested in estimating accurately the relative abundance of the present organisms. We can assess the abundance estimates produced by the methods by using error measures such as the relative root mean square error, RRMSE and the average relative error, AVGRE. For metaMix, we use the relative abundance estimates from the 25 runs. For all methods, when the exact strain was not identified but the correct species or genus was, we used this abundance.

$$\textbf{RRMSE} = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(\frac{|w_j - t_j|}{t_j})^2} \tag{4.1}$$

$$\textbf{AVGRE} = \frac{1}{K}\sum_{j=1}^{K}(\frac{|w_j - t_j|}{t_j}) \tag{4.2}$$

where t_j is the true abundance of species j and w_j the estimated abundance.

metaMix produces the most accurate abundance estimates and the results are summarized in Table 4.2.

4.3.3 Importance of read support parameter

We then assessed the importance of the read support parameter `r` on the output of metaMix. We ran metaMix on the benchmark simHC FAMeS dataset with `r` = {10, 20, 30, 50} reads, 25 runs for each (Table 4.3). We observe that as `r` decreases, a few more related strains from the reference database that are not in the community are retained in the output. As `r` increases two similar strains are merged into one.

We compared these results with the output of Pathoscope and MEGAN. None of these methods have a read support parameter serving the same purpose as in metaMix, so we tuned the most relevant parameters in these tools. Pathoscope has a thetaPrior parameter that enforces a unique read penalty. This parameter represents the read pseu-

Table 4.2: Number of species identified for the FAMeS simLC and simMC datasets, as well as sensitivity, specificity and abundance estimates error measures RRMSE and AVGRE. The metaMix results are based on 25 runs.

	metaMix	**Pathoscope**	**MEGAN**
		simHC	
Number of Species	116	165	232
Sensitivity	99.96	99.1	100
Specificity	99.8	97.7	95.0
RRMSE	16.9	36.6	35.9
AVGRE	8.3	29.7	18
		simLC	
Number of Species	114	147	208
Sensitivity	98.8	97.3	100
Specificity	99.8	98.4	95.9
RRMSE	21.1	185.6	32
AVGRE	8.9	53.3	16.1
		simMC	
Number of Species	115	144	208
Sensitivity	98.5	98.2	100
Specificity	99.8	98.6	95.9
RRMSE	29.6	152.7	31.9
AVGRE	12.9	49.2	19.3

docounts for the non-unique matches and the default setting is zero which allows for non informative priors. Using the default setting Pathoscope identifies 47 taxa. When thetaP is in (1,7) it identifies 22 taxa, while with thetaP>7 it identifies 165. With this latter setting which is the one we chose for the comparison, Pathoscope behaves as a standard mixture model.

MEGAN has a “Min Support” parameter which sets a threshold for the number of reads that must be assigned to a taxon so that it appears in the result. Any read assigned to a taxon not having the required support is pushed up the taxonomy until a taxon is found that has sufficient support. We used Min support = $\{10, 20, 30, 50\}$ reads. The respective number of taxa in the summary files were 250, 243, 236, 232.

We then also applied a post-run read count threshold to both methods’ output summary. We set the threshold for 10,20,30,50 reads respectively, disregarding taxa that have less than that number of reads assigned to them. In all instances metaMix produces a community profile closer to the real one, along with a better balance of sensitivity and specificity compared to the other two methods (Table 4.3). Pathoscope finds $\sim$ 15 more false positives while MEGAN $\sim$ 40 more compared to metaMix at the same read support level, except for the lowest `r`=10 where metaMix and Pathoscope achieve the same specificity. We also report further results using different posterior probability cutoffs for the different `r` settings.

Table 4.3: simHC FAMeS dataset: number of species (sd in parenthesis), sensitivity and specificity by metaMix (25 runs), Pathoscope and MEGAN, as a function of the min. number of reads required for each species to appear in the output. metaMix: `r`=$\{10, 20, 30, 50\}$ reads, Pathoscope: thetaPrior$>$ 7+ post-run threshold =$\{10, 20, 30, 50\}$ reads, MEGAN: "Min Support" + post-run threshold =$\{10, 20, 30, 50\}$ reads.

Read Support	metaMix	Pathoscope	MEGAN
50	114 (0.9)	131	147
Sensitivity - Specificity	99.1 - 99.9	98.2 - 99.1	100 - 98.5
30	116 (0.95)	131	156
Sensitivity - Specificity	99.96 - 99.8	98.2 - 99.1	100 - 98.2
20	124 (1.65)	141	166
Sensitivity - Specificity	100 - 99.5	98.2 - 98.7	100 - 98
10	155 (1.9)	155	188
Sensitivity - Specificity	100 - 98.2	99.1 - 98.2	100 - 97.4

4.3.4 simMC and simLC communities

The FAMeS project includes two additional mock communities that consist of the same 113 species as simHC, but they differ in their relative abundances setup: in simLC (low complexity) there is one dominant species or a few more in simMC (medium complexity). We ran metaMix 25 times for both, changing the random seed. These 2 datasets turned out to be more challenging for all three methods, missing or merging together some similar related strains. metaMix outperforms Pathoscope and MEGAN in terms of producing a parsimonious community profile and having the best sensitivity and specificity trade-off (Table 4.2).

For the simMC and simLC communities we also ran metaMix for different posterior probability cutoffs (0.5-0.9) and different read support values (`r`=$\{10,20,30\}$). We present the results in Tables 4.4 and 4.5. Naturally, as we allow species with lower posterior probabilities in the results, the sensitivity increases and the specificity decreases.

Changing the read support value and comparing with MEGAN and Pathoscope, we observe the same pattern as for simHC: metaMix has the best balance of specificity and sensitivity between the three methods.

Table 4.4: simLC, simMC: Number of species detected by metaMix as well as sensitivity, specificity, AVGRE, RRMSE for metaMix at various posterior probability cutoffs. The results are average values based on 25 runs.

Cutoff	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**
simLC					
Sensitivity (mean)	98.32	**98.82**	99.00	99.07	99.11
Sensitivity (sd)	0.0083	**0.0050**	0.0030	0.0018	0
Specificity (mean)	99.89	**99.85**	99.82	99.78	99.75
Specificity (sd)	0.0004	**0.0003**	0.0004	0.0004	0.0002
# Species - median	113	**114**	115	116	117
# Species - sd	1.4	**1.1**	1.0	0.9	0.4
rRMSE	21.1	**21.0**	21.1	21.3	21.6
AVGRE	8.9	**8.8**	8.9	8.9	9.2
simMC					
Sensitivity (mean)	97.96	**98.46**	98.79	98.93	99.11
Sensitivity (sd)	0.0061	**0.0048**	0.0044	0.0036	0
Specificity (mean)	99.83	**99.77**	99.71	99.66	99.63
Specificity (sd)	0.0005	**0.0004**	0.0004	0.0004	0.0002
# Species - median	114	**115**	118	119	120
# Species - s.d	1.17	**1.07**	1.08	1.08	0.48
RRMSE	29.98	**29.93**	30.05	30.11	29.96
AVGRE	13.05	**13.18**	13.31	13.37	13.26

4.3.5 simHC - assembled data

The results we have reported in the main text far are based on unassembled simHC FAMeS data. We subsequently wanted to compare the performance of metaMix on the same dataset, doing first an assembly step. We used Velvet with a high kmer value ($k = 89$) in order to obtain high quality contigs. This resulted in 733 contigs made up by 2,403 reads, i.e approximately 2% of the total reads were contributing to contigs.

Table 4.5: simLC, simMC FAMeS datasets: number of species detected as well as sensitivity and specificity of metaMix, Pathoscope and MEGAN, as a function of the minimum number of reads required for each species to appear in the output. For metaMix that is `r`=$\{10, 20, 30\}$ reads, for Pathoscope thetaPrior$>$ 7$+$ post-run threshold =$\{10, 20, 30\}$ reads, for MEGAN "Min Support" + post-run threshold =$\{10, 20, 30\}$ reads.

	metaMix	**Pathoscope**	**MEGAN**
		simLC	
r30	114 (1.09)	126	142
Sensitivity- Specificity	98.82 99.84	97.32 - 99.27	100 - 98.71
r20	116 (1.17)	127	147
Sensitivity- Specificity	98.89 99.76	97.32 - 99.22	100 - 98.5
r10	133 (1.22)	131	157
Sensitivity- Specificity	100 99.11	97.3 - 99.05	100 - 98
		simMC	
r30	115 (0.69)	126	141
Sensitivity- Specificity	98.46 99.77	98.21 - 99.35	99.1 - 98.8
r20	117 (1.1)	126	145
Sensitivity- Specificity	98.21 99.67	98.21 - 99.35	99.1 - 98.6
r10	144 (2.3)	130	158
Sensitivity- Specificity	99.46 98.56	98.2 - 99.18	99.1 - 98

We then annotated contigs and unassembled reads with BLASTn and applied metaMix with default parameters. We find all members of the metagenomic community and one false positive (based on one run: sensitivity=100, specificity=99.96). The estimates for relative abundance were also close to the true values ($RRMSE = 16.9$ and $AVGRE = 8.2$). We therefore observe that the metaMix results are very similar whether we choose to include or forego an assembly step, with the resulting community profile very close to the true one.

4.3.6 Comparison of IS - Defensive Sampling - MLE

We compared the performance of metaMix on the same simHC FAMeS dataset, using Importance Sampling and Defensive Importance Sampling (95% samples produced by posterior approximating g and 5% by π) for the marginal likelihood estimation as well as using the MLE approximation. In the above section, we discussed the results using the latter option. For the IS and the defensive IS of the marginal likelihood, 1,000 samples were drawn from the importance distribution for each model considered, i.e at each MCMC iteration .

The resulting species profiles can be seen in Table 4.6. We ran all three versions of metaMix for 1,000 MCMC iterations in order to obtain results within 24 hours. All the approaches produced almost identical results, in terms of species identified and abundance estimates, with the defensive IS performing slightly better in terms of abundance estimation accuracy. However the MLE approximation method was ∼13x times faster than the other two, reducing the time required from ∼19 hours to 90 minutes.

Table 4.6: FAMeS simHC - comparing the effect of different marginal likelihood estimation methods on metaMix performance: species profiling, accuracy of abundance estimation and computational time.

	Importance Sampling	**Defensive Sampling**	**MLE approximation**
Number of species	116	116	116
$\boldsymbol{w}$ estimate - rRMSE	17	16.8	17.1
$\boldsymbol{w}$ estimate - AVGRE	8.5	8.3	8.6
Computational time (hours)	18.6	18.6	1.5

4.3.7 Benchmarking conclusions

With the benchmark datasets used, metaMix provides a good balance of sensitivity and specificity, outperforming MEGAN and Pathoscope. A consequence of the increased accuracy is that metaMix produces better estimates for the relative abundances of the species in the mixture. The method can deal with either unassembled reads or assembled contigs or both, allowing for flexibility of choice for the bioinformatics preprocessing.

The FAMeS datasets are complex and distinct from the typical human clinical samples we have worked with, where we wish to detect the presence of viral infectious agents. In these we typically have mixtures of eukaryote (human), bacterial and viral sequences and we require the method to be sensitive enough to be able to detect viral traces. Additionally the samples we work with are normally from sterile sites, therefore we do not expect a large number of organisms. At the time of testing metaMix there were no mock datasets with viral infections so that we could benchmark the methods in a scenario more similar to the one we have developed metaMix for. This is the reason we extended the methods comparison to the following chapter 5 on the sequencing data from human clinical samples with viral infections. These viral infections were unknown prior to the metaMix analysis of the data, however they were subsequently confirmed using other molecular methods. Despite the differences, the FAMeS dataset are essential to use as benchmark for examining the performance of the methods in varyingly complex communities, where there are also some closely related strains in the sample.

Chapter 5

Pathogen identification in clinical samples

We discussed in the introduction the potential that deep-sequencing technology has for detecting pathogens in clinical samples. Diagnostic tools based on high-throughput sequencing can offer valuable insight, especially for patients suffering from infectious diseases. An important feature of this methodology is that it does not make prior assumptions about the type of pathogen, but has the potential to detect DNA or RNA from all species.

In this chapter we discuss the use of metaMix as a diagnostic tool in clinical cases where the pathogen could not be identified with conventional testing. UCL is the primary research partner of the Great Ormond Street Hospital for Children (GOSH) where academics and clinicians collaborate towards diagnosing and treating childhood disease. Our collaboration is with Professor Judith Breuer at the Division of Infection & Immunity in UCL and who is also a consultant Virologist at GOSH. Following the successful diagnosis of the first case (discussed below), Prof. Breuer has now adopted the use of RNA-seq as part of routine testing for all undiagnosed cases. In all cases a plausible infectious pathogen was identified, using metaMix for read interpretation and confirmed with laboratory methods. These are PCR assays, discussed in chapter 1 and immunohistochemistry methods. All bioinformatics and statistical analysis of the generated data were undertaken by me, while the laboratory work to confirm the metaMix finding was done by Julianne Brown, PhD candidate in the Breuer group.

The PCR assays provide information on viral load and results are obtained using separate RNA extractions from the clinical sample. Cycle threshold (CT) is a relative measure of the concentration of target in the PCR reaction, with values representing the cycle number at which amplification was detected. CT has an inverse relationship with viral titer, with a small value indicating high titer. In clinical practice, CT>38 may be considered ambiguous but in the context of other positive results, is considered a true positive. Immunohistochemistry (IHC) is another technique used for confirmation purposes, designed to detect the presence of abnormal cells. IHC is based on the interaction of target antigens with specific antibodies tagged with a visible label. This allows the visualization of specific cellular components within cells that act as markers.

Here we discuss the first two cases of two immunosuppressed patients who developed an infectious disease, more specifically encephalitis. We first give some background on encephalitis and then we present and discuss the results for each case.

5.1 Encephalitis background

Encephalitis is a complex and potentially devastating neurological syndrome, characterized by inflammation in the brain and associated with brain dysfunction (Thompson *et al.*, 2012). In the majority of encephalitis cases direct viral infection, that is the virus crossing the blood-brain barrier, is thought to be the cause (Granerod *et al.*, 2010). However in over half of the cases a causative agent is not found (Glaser *et al.*, 2006). The gold standard diagnostic testing for encephalitis is polymerase chain reaction (PCR), a method suitable for determining the amount of a target sequence that is present in a sample. PCR requires prior knowledge of potential infectious agents and can therefore be narrow in scope. Different treatment options for infectious and non-infectious cases render understanding the pathology of encephalitis of crucial clinical importance. The most commonly identified causes of encephalitis are the herpes simplex virus (HSV) and the varicella zoster virus (VZV), however immunocompromised patients are more likely to get encephalitis from a greater variety of viruses (Thompson *et al.*, 2012). Deep sequencing of clinical samples has increasingly been used as a diagnostic tool (Barzon *et al.*, 2013) (Handley *et al.*, 2012), specifically in cases where traditional techniques fail to unveil the disease-associate pathogens (Chiu, 2013) (Wilson *et al.*, 2014).

5.2 Clinical Case 1

The first case was an immunosuppressed 18-month old boy with a case of encephalitis for which traditional tests could not find the causative agent. Using RNA-seq from a brain biopsy, we identified an astrovirus, highly divergent from human astrovirus genotypes typically associated with diarroea, but closely related to an astrovirus that has been once before implicated in a fatal encephalitis in an immunocompromised patient (Quan *et al.*, 2010). The work undertaken is published (Brown *et al.*, 2015), our clinical collaborators having obtained written consent for publication from the family.

5.2.1 Case Report

The patient had Cartilage Hair Hypoplasia (CHH) and associated immunodeficiency. CHH is a rare autosomal recessive disorder characterised by limited bone growth. The patient underwent an uncomplicated peripheral blood stem cell transplant in GOSH in 2013, however two weeks later he became acutely unwell with irritability, dystonia and reduced consciousness. The differential diagnosis concluded this was a case of encephalitis of infectious aetiology. An extensive PCR panel for 18 different viruses, including the Human Astrovirus, on cerebrospinal fluid (CSF) after the onset of symptoms at three different time intervals (two days, two weeks and one month after the onset) was negative for all the tested pathogens. Brain biopsy performed eight weeks after the neurological deterioration was also negative for these pathogens, as well as fungal and bacterial ribosomal RNA gene PCRs. It was then decided to perform deep sequencing on the brain biopsy sample. The pathogen detection from the RNA-seq analysis informed subsequent patient management. Unfortunately the child died nine months after the transplantation, in the context of ongoing neurological impairment with recurrent respiratory and gastrointestinal complications.

5.2.2 Data and preprocessing

Total RNA was purified from the biopsy and polyA RNA was separated for sequencing library preparation. Rapid RNA-Seq was performed on an Illumina Miseq. The Illumina MiSeq instrument generated 20 million paired-end reads.

I analysed the raw data using the bioinformatics pipeline and reference database described in Chapter 1. Non-clonal and good quality reads made up to 90% of the dataset, however only 4% of the reads, i.e $\sim 75{,}000$ were non-human. Based on the

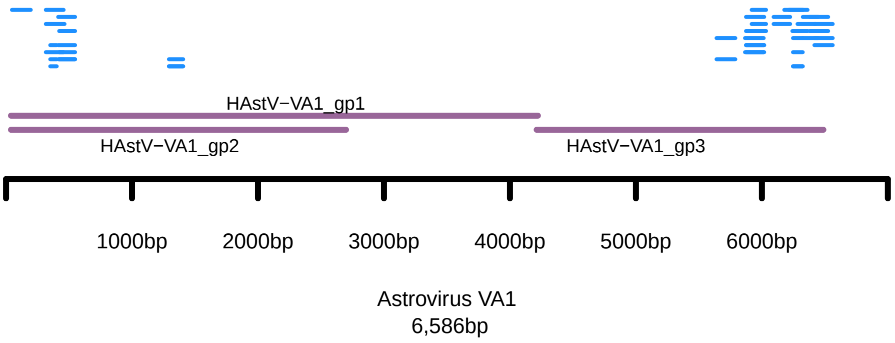


Figure 5.1: Clinical Case 1 - novel virus.
The reads (blue lines) assigned by metaMix to Astrovirus VA1, aligned to its genome. The purple lines represent the genes of the virus.

BLASTx output there were 1,298 potential species.

5.2.3 metaMix results

Following the initial bioinformatics processing, we used metaMix for species identification and abundance estimation. The resulting species profile is shown in Table 5.1; the 13 metaMix entries correspond to 10 species. The most abundant organism was the ΦX174 bacteriophage, which is routinely used for deep-sequencing quality control. The bacteriophage could have been filtered out prior to community profiling. However this was the first clinical case metaMix was used on and its role as positive control was useful.

More interestingly, we identified an astrovirus. Five short assembled contigs (44 reads) with length ranging from 167bp to 471bp and two non-assembled reads were assigned to the *Astrovirus VA1*. The posterior probability of the virus being present in the sample was one (Figure 5.1). The individual read classification probabilities were high (plotted in Figure 5.2), indicating that the reads were unambiguously assigned. metaMix also identified a number of bacteria supported by a few reads. These are either known human skin associated contaminants or laboratory reagent or extraction kit contaminants (Salter *et al.*, 2014). The analysis completed in 29 minutes.

The contig sequences generated in the bioinformatics step were used to design a PCR assay specific to the astrovirus identified in this study. This confirmed the astrovirus at a high viral titer in the brain biopsy, as indicated by the CT value at 25. The PCR assay that was designed based on the assembled contigs (AstV-contig) was used

Table 5.1: Clinical Case 1 - metaMix summary profile consisting of twelve taxa.

Taxon Identifier	Scientific Name	Assigned Reads	Posterior Probability
374840	*Enterobacteria phage phiX174 sensu lato*	60449	1
NA	*unknown*	10254	1
9606	*Homo sapiens*	216	1
28090	*Acinetobacter lwoffii*	89	1
469	*Acinetobacter*	78	0.99
13690	*Sphingobium yanoikuyae*	60	0.99
645687	*Astrovirus VA1*	46	1
133448	*Citrobacter youngae*	37	0.91
199310	*Escherichia coli CFT073*	33	1
56946	*Afipia broomeae*	28	1
618	*Serratia odorifera*	23	0.92
409438	*Escherichia coli SE11*	21	0.98
1747	*Propionibacterium acnes*	13	0.97

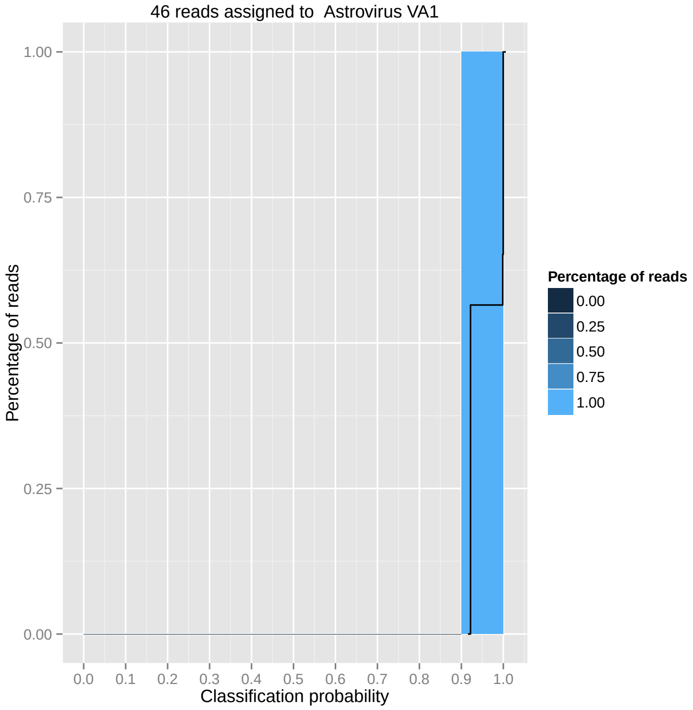


Figure 5.2: Clinical Case 1 - Classification probabilities for the detected Astrovirus.

to retrospectively test all stored samples from the patient during his hospital admission. Even though the highest titers of astroviral RNA were found in the brain tissue, viral RNA was also detected in the CSF as well as in stool and serum samples, the latter confirming viremic spread.

Sections of the brain biopsy were stained using AstV-specific antibody for the capsid protein and the results revealed extensive staining of cell bodies. The positive immunohistochemistry for the capsid protein provided further confirmation of replication competent virus in the brain.

Additionally, the full viral genome sequence was generated using overlapping PCR and subsequent Sanger sequencing. Phylogenetic analysis of the RNA-dependent RNA polymerase nucleotide sequences (shown in Figure 5.3, taken from our paper (Brown *et al.*, 2015) and generated by Julianne Brown) showed that the virus identified in the brain clustered with a group of viruses more closely related to animal astroviruses than the human HAstV 1-8 (less than 47% pairwise similarity) primarily associated with gastroenteritis. The novel virus was termed HAstV-VA1/HMO-C-UK1 and it exhibited 98%, 97% and 95% pairwise similarity with the VA1, HMO-C, SG strains respectively. The latter has once before been reported to cause fatal encephalitis to a 15 year old boy with X-linked agammaglobulinemia (Quan *et al.*, 2010). This further points towards a pathogenic role for this virus group, likely to have been previously under-recognised in immunocompromised patients.

Finally, to investigate the possible source of infection, all stored specimens from patients on the same ward as our patient were tested by the AstV-contig PCR assay. This included eighteen samples (stool and urine) from nine patients that were collected from two weeks before to two weeks after our patient's first positive AstV-contig PCR result, There was a single patient who had a positive stool sample. There was insufficient residual sample for repeat testing by quantitative PCR or for sequencing, therefore the positive result could not be confirmed. No other stored samples from this second patient were positive.

5.2.4 Comparing metaMix results to other methods

We also analysed the dataset using both Pathoscope and MEGAN to compare the performance of the three tools in a clinical sample where the viral load is low.

Pathoscope

Pathoscope identified 22 taxa, corresponding to 15 species and some genera or families (Appendix Table B.3). It also assigned all 46 reads to the *Astrovirus VA1*. Almost all the species identified from metaMix were identified by Pathoscope, with an additional 9 taxa supported by few reads. As the method can only properly work with unassembled

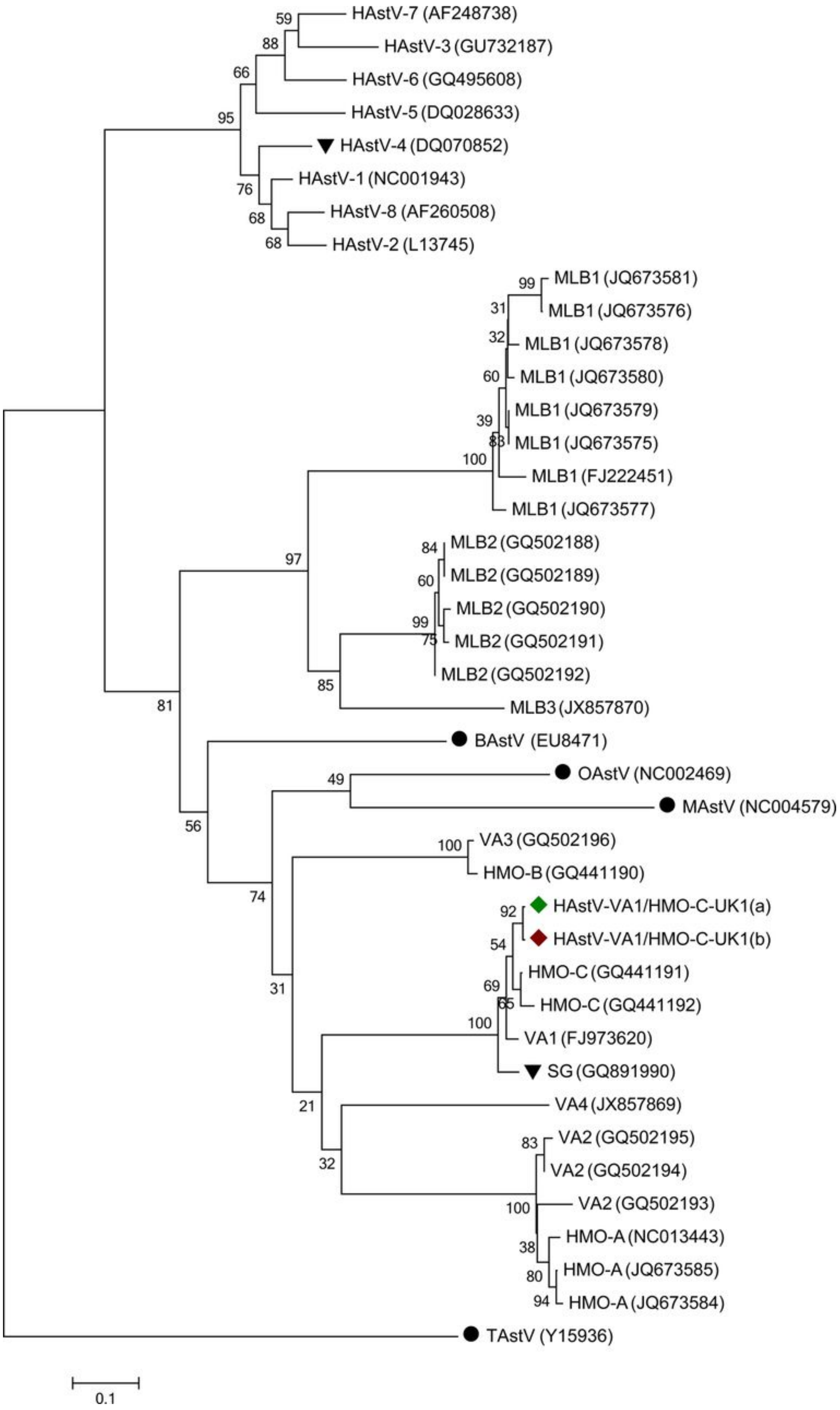


Figure 5.3: Clinical Case 1 - Phylogenetic analysis based on the RdRp gene. The astrovirus (AstV) identified in the brain of the patient is denoted by a green rhombus (HAstV-VA1/HMO-C-UK1(a)) while the AstV identified in the second patient's stool by a red rhombus. The black inverted triangles indicate AstV species previously reported in patients with neurological disease. The black circles indicate sequences of animal origin. Abbreviations: BAstV, bat astrovirus; MAstV, mink astrovirus; OAstV, ovine astrovirus; TAstV, turkey astrovirus. Scale bar represents the number of base substitutions per site.

sequence data, an extra BLASTx similarity step had to be performed for the 91,516 reads that had contributed to the 679 assembled contigs. Pathoscope produced the results in less than one minute.

MEGAN

MEGAN identified 19 taxa and did not detect the astrovirus signal. We modified the minimum read support parameter from 50 reads to 10 to increase sensitivity. MEGAN then identified 25 taxa, including the *Astrovirus VA1*. The remaining 24 were mostly genera, relevant to the species detected by metaMix and Pathoscope. MEGAN produced the results in less than one minute.

Kraken

We also used the Kraken App for the Illumina BaseSpace. BaseSpace (`http://basespace.illumina.com`) is the Illumina genomics computing environment for high-throughput sequencing data analysis and management, available as a cloud solution. There are several Applications (Apps) that the BaseSpace user can launch to analyse data online. Kraken (Wood and Salzberg, 2014) mentioned previously in chapter 1, assigns taxonomic labels to reads using exact alignments of k-mers and the LCA approach. The BaseSpace uses version 0.10.4-beta and the MiniKraken reference database which contains bacteria, archaea and viruses. The raw data were used along with the option for host read filtering.

Kraken classified 78% of the reads as human. Of the non-host reads 72% subsequently remained unclassified, while it classified the other 28% in 907 taxa. Among these there were 108 viruses including three astroviruses. There was no parameter tuning option available on the BaseSpace version of Kraken. It is plausible that customisation, i.e setting different k-mer values, of the command line version of the tool may correct the poor specificity to some degree. However assessed on default settings and compared to the other three methods discussed in the previous sections, it manifests worst specificity.

PhyloSift

Finally we used PhyloSift, which employes a phylogenetic approach. PhyloSift uses a core set of genes and identifies the phylogenetic relationship of the read sequences to the database sequences. We used the host filtered data as an input to PhyloSift. The summary with the taxa present in the sample consists of 1217 taxa, including 219 viral taxa and 5 astroviruses. A filtered version of the summary which contains only candidate sequences where the placement probability is greater than 0.9, has removed a substantial number of erroneous taxon assignments, however it still consists of 191 taxa, 94 viruses and 4 astroviruses. Along with Kraken, they exhibited the worst specificity among the tested methods. As mentioned in Chapter 1, PhyloSift is designed to infer the large phylogenetic framework rather than act as a taxonomic classification method and therefore the comparison is a little unfair and the results perhaps not surprising. However we included it for the sake of completeness and to highlight that when accurate and specific detection is of interest, a similarity Bayesian mixture model is the best choice.

5.3 Clinical Case 2 - Coronavirus

The potential for deep sequencing to detect nucleic acids from a broad range of species and the success in the astrovirus case prompted us to use RNA-sequencing of a brain biopsy from an immunosuppressed 10-month-old boy with symptoms of viral encephalitis in whom conventional diagnostic PCRs were negative. We identified a newly emerged strain of human coronavirus OC43 as the most likely causative agent. In recent years Coronaviruses have emerged as increasingly important pathogens in serious human infections. However this is the first instance to confirm the associations between coronavirus and central nervous system (CNS) disease in humans, long hypothesized based on observations from mouse models. The work undertaken is under review (Morfopoulou *et al.*, Identification by deep-sequencing of a novel human coronavirus OC43 strain in an infant with severe combined immunodeficiency and fatal encephalitis), our clinical collaborators having obtained written consent for publication from the patient's family.

5.3.1 Case Report

A male infant suffered recurrent respiratory infections and failure to thrive from 4 months of age. He was diagnosed at 7 months with severe combined immunodeficiency (SCID) and at 9 months while waiting for a bone marrow donor, deteriorated with the onset of neurological symptoms. Magnetic resonance imaging (MRI) revealed

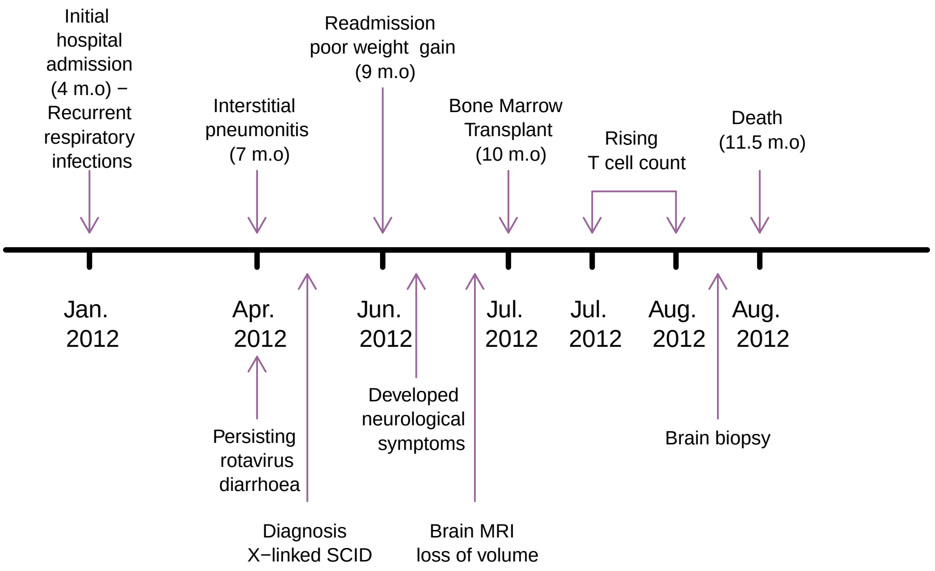

Initial hospital admission (4 m.o) – Recurrent respiratory infections
Interstitial pneumonitis (7 m.o)
Readmission poor weight gain (9 m.o)
Bone Marrow Transplant (10 m.o)
Rising T cell count
Death (11.5 m.o)
Jan. 2012
Apr. 2012
Jun. 2012
Jul. 2012
Jul. 2012
Aug. 2012
Aug. 2012
Persisting rotavirus diarrhoea
Diagnosis X−linked SCID
Developed neurological symptoms
Brain MRI loss of volume
Brain biopsy

Figure 5.4: Clinical timeline for case 2. Arrows mark important events over the course of the illness.

viral encephalitis. Apart from rotavirus in stool, an extensive panel of PCR for viruses and bacteria on cerebrospinal fluid (CSF), blood, stool, urine and nasopharyngeal aspirate did not identify a pathogen. At 10 months he underwent a cord blood transplant, however he continued to deteriorate. A brain biopsy was taken at 11 months and the patient died 1.5 months post-transplant. A schematic of the clinical timeline can be seen in Figure 5.4.

5.3.2 Data and preprocessing

Total RNA was purified from a frozen brain biopsy and polyA RNA separated for sequencing library preparation. 64 million RNA-Seq 80bp paired-end reads were obtained using the HiSeq Illumina 2500 instrument. We processed the raw data using the bioinformatic pipeline described in Chapter 1. 1.4 Million reads were identified as non human, accounting for 2% of the raw data. These were then annotated with BLASTx against the custom protein database mentioned. The BLASTx output contained matches to 3,150 potential species.

5.3.3 metaMix results

The resulting species profile consisted of 7 species 5.2. The method ran in 4.7 hours. The vast majority of the reads (∼one million) were assigned to Human coronavirus

OC43, with 67K reads also matching a different human coronavirus. We see that despite the initial host filtering, a few thousand reads were annotated as human on the protein level. A few hundred reads matched environmental bacteria and known laboratory contaminants.

Table 5.2: Clinical Case 2 - metaMix summary profile consisting of eight taxa.

Taxon identifier	**Scientific Name**	**Assigned Reads**	**Posterior Probability**
31631	*Human coronavirus OC43*	997453	1
unknown	*unknown*	170535	1
627439	*Human enteric coronavirus strain 4408*	67118	1
9606	*Homo sapiens*	23676	1
10090	*Mus musculus*	2301	1
47229	*Massilia timonae*	240	0.99
85698	*Achromobacter xylosoxidans*	127	0.9
72556	*Achromobacter piechaudii*	64	0.9
56946	*Afipia broomeae*	61	0.97

The summary contains two different human coronaviruses, however upon further study both entries indicate the sole presence of an HCoV-OC43 strain. GenBank currently lists 58 HCoV-OC43 strains, but our database contains only the RefSeq sequence, which is the laboratory prototype strain first isolated in 1967 (Vijgen *et al.*, 2005b). These other strains feature genomic insertions and deletions in reference to the prototype strain (Vijgen *et al.*, 2005a). Some of the dissimilarities are present in the second coronavirus species we identified, as well as in the virus in our sample. Therefore metaMix results suggest that consideration of the different HCoV-OC43 strains could reveal the closest one to the virus in the sample.

This is further supported by the cumulative histogram plot of the classification probabilities for each species in the summary profile (Figure 5.5). There is a very clean signal for HCoV-OC43 and we see that the vast majority of the reads assigned to it have classification probabilities greater than 0.9. On the other hand only 25% of the reads assigned to the 4408 species have high classification probabilities while the remaining have very poor ones. Therefore, while a few thousand reads match better to the 4408 proteins rather than HCoV-OC43 ones, the most plausible explanation is that the virus in the brain looks more like the HCoV-OC43 but in some regions shares greater similarity with the second coronavirus species. There are a number of reasons why a mixed infection is unlikely: first, even though the occurrence of multiple virus infections in immunodeficient patients is not uncommon, mixed viral infections that occur at the same site and involve viruses with similar expressions of disease are less frequent (Waner, 1994). When coinfection of the same specimen does occur, it usually involves viruses from different families (such as rhinoviruses, coronaviruses, influenza A viruses). Finally, the *Human enteric coronavirus strain 4408* was originally isolated from the stool of a child from a rural area with acute diarrhoea (Zhang *et al.*, 1994) and shows greater relatedness to bovine coronaviruses than human coronaviruses. As opposed to HCoV-OC43, this is not a common, circulating human pathogen.

We mentioned previously issues with the choice of reference database and how depending on the database the answer may differ. Our choice included the viral RefSeq database, a curated and well annotated collection of sequences. Its limited size renders it a suitable choice for annotating sequences prior to community profiling due to reasonable computation times. However depending on the question of interest, additional analyses using a more inclusive database may be undertaken for greater taxonomic resolution, as was done in this case and described in the next section.

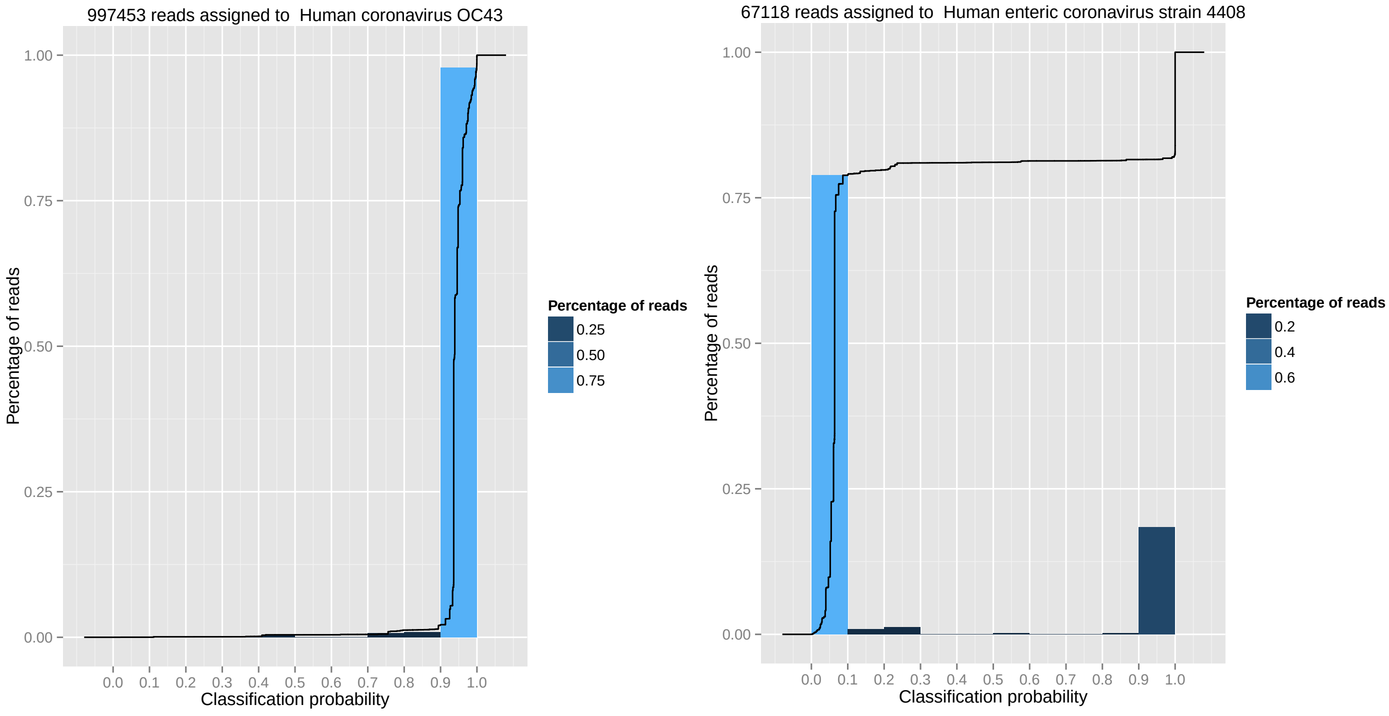


Figure 5.5: Clinical Case 2 - Classification probabilities for the 2 coronaviruses in the results.

The presence of HCOV-OC43 was confirmed by real-time PCR performed in brain tissue with a CT value of 24, as well as positive IHC staining in neurons.

5.3.4 Phylogenetic Analysis

We then wanted to see how the viral sequence detected in the brain biopsy compared to other HCoV-OC43 sequences in GenBank. A consensus sequence was extracted from the mapped reads, using Samtools (Li *et al.*, 2009a) (version 0.1.19) and the de novo assembled sequence by Velvet as a template. Multiple sequence analysis was performed using ClustalO (Sievers *et al.*, 2011) between the consensus sequence of the virus from the sample and selected sequences from GenBank. A full-genome phylogenetic maximum likelihood tree was estimated by PhyML (Guindon *et al.*, 2010) using the multiplatform interface SeaView (Gouy *et al.*, 2010) (version 4.3.1). Details for the parameter settings of PhyML can be found in Appendix C.

Different studies support that HCoV-OC43 sequences can be classified in distinct classes (Vijgen *et al.*, 2005a) or genotypes (Lau *et al.*, 2011). In the most recent study (Zhang *et al.*, 2014) five potential genotypes are discussed, the oldest one being the original laboratory prototype (genotype A) and the most recently identified one being E. Genotype E has been discussed solely in the context of respiratory tract infections as it was identified in late 2014, despite being thought to have emerged in 2010.

We chose five sequences from GenBank (NC_005147, AY903459, JN129834, AY903460, KL198610), one to represent each genotype. The bovine coronavirus sequence (NC_003045) was used as an outgroup. The estimated full-genome phylogenetic tree reveals a clear similarity of the consensus viral sequence to the novel genotype E as seen in Figure 5.6. This is the most recent one, thought to be a recombinant from B, C and D (Zhang *et al.*, 2014) . We had additionally confirmed this was the case for our detected consensus viral sequence using bootscanning, a method for anaysis of viral recombination (Lole *et al.*, 1999). The main idea is that a potentially recombinant sequence is compared to a set of plausible parental sequences. We produce a multiple sequence alignment which is subsequently broken into sliding windows. Phylogenetic trees with bootstrap support are built for each window and the bootstrap value is plotted along the genome of interest. Figure 5.7 demonstrates the genetic distances of the detected HCoV-OC43 to genotypes B, C and D accross its genome.

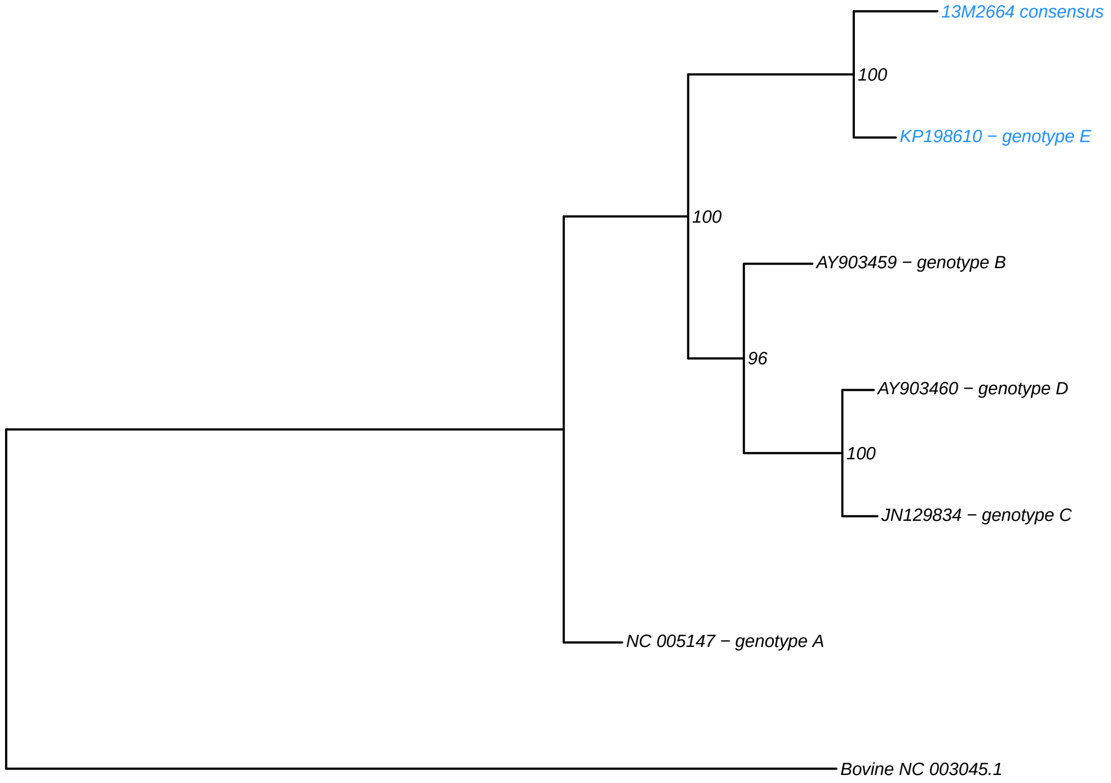


Figure 5.6: Clinical Case 2 - Full genome phylogenetic tree for the viral sequence identified in this study (13M2664 consensus in blue font) and the other 5 genotypes of coronavirus OC43. A, B, C, D and E represent OC43 genotypes.

The bootscanning analysis may appear redundant, however it was carried out when attempts at reconstructing the phylogenetic tree resulted in the brain virus not clustering with any genomes in the tree. This was due to the fact that the genotype E genome was not published in GenBank until January of 2015 (Zhang *et al.*, 2014) and thus, this sequence was not included in our initial analyses. Without any closely related sequences in the public databases, we wanted to understand how this virus related to the other genotypes and therefore performed this analysis.

5.3.5 Variant Detection

The large number of short reads mapping to the HCoV-OC43 genome enabled us to confidently identify single nucleotide polymorphisms (SNPs) as compared with the published genotype E of the HCoV-OC43 strain. We aligned all QC short reads to the genotype E genome (genbank identifier: KP198610) using novoalign, resulting in

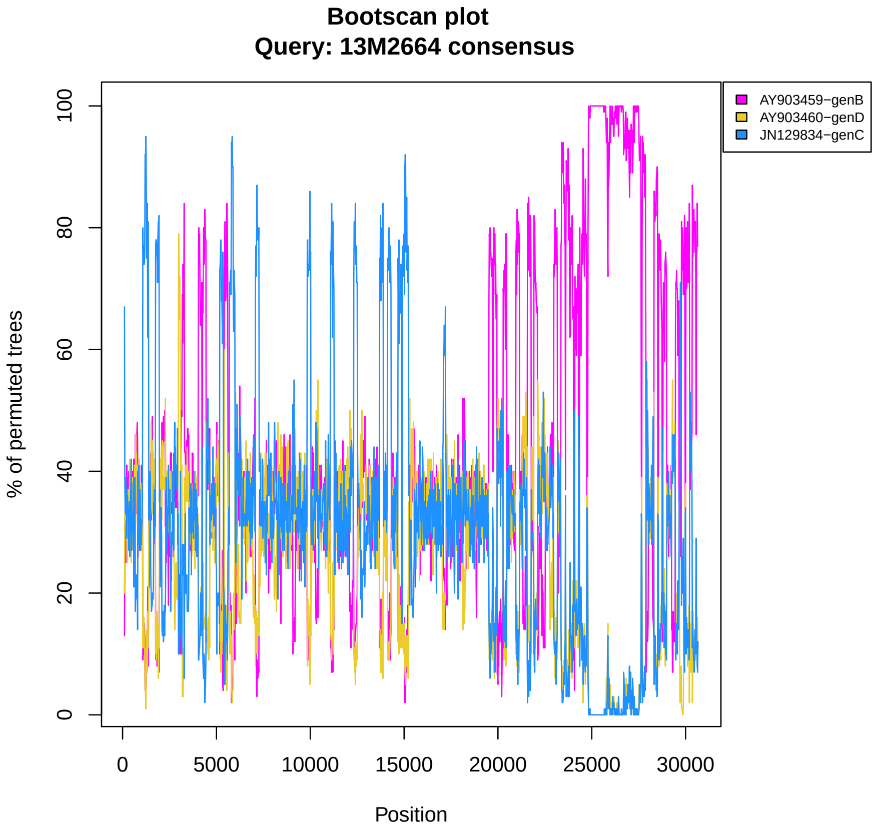


Figure 5.7: Clinical Case 2 - Recombination bootscan plot analysis of the viral sequence identified in this study (13M2664 consensus used as a query) compared to reference strains in genotypes B, C and D. The analysis used a sliding window of 200 bp and 100 bootstraps.

1,175,294 mapped reads. We identified 102 variants (quality>= 30 and depth>= 20) using Samtools (Li *et al.*, 2009a) (version 0.1.19) (Figure 5.8). From these 47 were missense variants, i.e resulting in a codon that codes for a different amino acid, based on annotation with SnpEff (Cingolani *et al.*, 2012). The full list of variants is in (Appendix Table B.5).

It is of interest to note that one third of the missense mutations occurred in the coding region of the S protein, a region with greater sequence variation compared to the rest (Arbour *et al.*, 1999). The S protein is the main viral protein involved in receptor recognition on the cell surface and its role in determining neurovirulence has been established in various studies (St-Jean *et al.*, 2004; Pierre J. Talbot and Jacomy, 2011; Desforges *et al.*, 2014). However none of the mutations we identified has been previously associated with increased neurovirulence (Favreau *et al.*, 2009; Jacomy *et al.*, 2006).

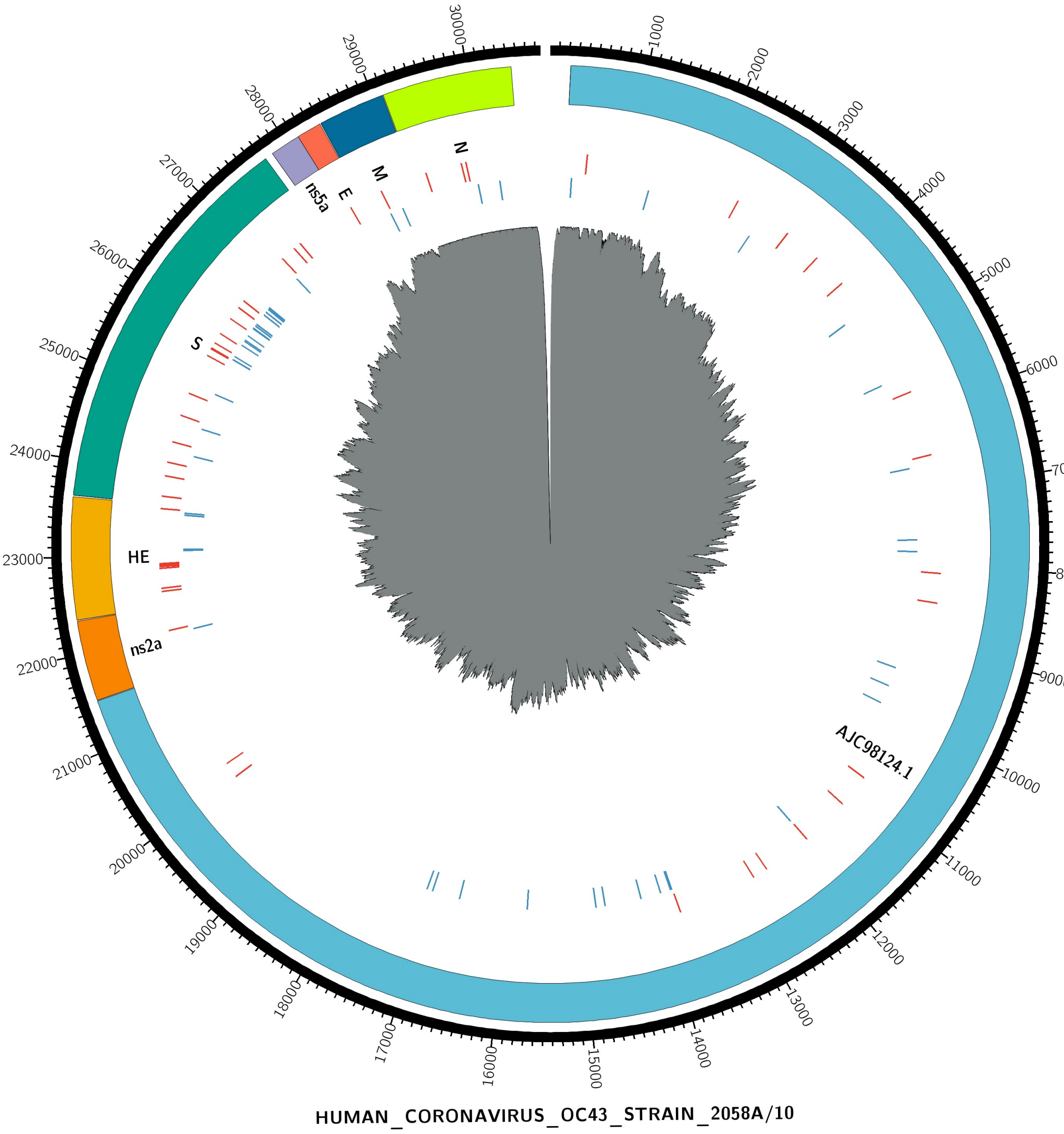


Figure 5.8: Clinical Case 2 - Circos plot displaying the genomic structure of human coronavirus OC43 strain 258A/10 (genotype E) and the variants called. Missense variants are in red, while synonymous in blue. The grey histogram in the middle is the read coverage (log10 values).

5.3.6 Discussion

Encephalitis is a severe neurological pathology, characterized by parenchymal inflammation of the brain, rare in the general population but more frequent in immunocompromised patients. The combination of a large number of pathogens known to cause viral encephalitis and the use of PCR for diagnostic testing contribute to the high frequency of unknown aetiology encephalitis cases. Here we diagnosed post-mortem an immunocompromised infant with acute encephalitis using deep sequencing on frozen brain biopsy. We identified high RNA levels of a novel HCoV-OC43 strain, confirmed by PCR and immunohistochemistry. This is one of the few times where direct evidence of HCoV RNA has been detected in the brain of a patient with encephalitis.

Coronaviruses (CoVs) are potentially lethal pathogens of the Coronaviridae family, a group of linear single-stranded enveloped RNA viruses, with the largest genome ($\sim$ 31 kb) among known RNA viruses. They are of great interest to human and animal health and are associated with a broad infection spectrum. CoVs are primarily recognised as respiratoric pathogens, involved in mild to serious lower respiratory tract infections (Perlman and Netland, 2009). They have been associated with a wide range of disorders such as pneumonia, encephalitis, hepatitis and enteritis (Pierre J. Talbot and Jacomy, 2011). Depending on the coronavirus type, they first interact with respiratory tract and mucous cells and can potentially spread to other tissues, including the central nervous system (Pierre J. Talbot and Jacomy, 2011; Desforges *et al.*, 2014).

Coronaviruses are divided into four groups, Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus. Betacoronaviruses include the human coronaviruses which are major causes of the common cold such as HCoV-OC43 and HCoV-HKU1, and can occasionally cause pneumonia. They also include SARS-CoV (Peiris *et al.*, 2003; Marra *et al.*, 2003) and MERS-CoV (Zaki *et al.*, 2012) that emerged in the last fifteen years and cause severe acute respiratory syndrome, with a high rate of mortality and morbidity, resulting in the revived interest of the scientific community in the species (Graham *et al.*, 2013) .

Three human CoVs OC43, 229E as well as SARS have been shown to have neuroinvasive and neurotropic properties (Arbour *et al.*, 1999, 2000; Xu *et al.*, 2005). In vivo studies in mice show that HCoV-OC43 is able to infect neuros and cause encephalitis (St-Jean *et al.*, 2004; Jacomy *et al.*, 2006) . The virus has also been shown to cause persistent infections in human neural cell lines (Favreau *et al.*, 2009). Interestingly, HCoV-OC43 RNA has been detected by PCR in human brains from multiple sclerosis (MS) patients and healthy subjects (Arbour *et al.*, 2000). A single case report identified HCoV-OC43 RNA in the CSF of a child with acute disseminated encephalomyelitis (Ann Yeh, Arlene Collins, Michael Cohen and Faden, 2004). However this is the first case in which HCoV-OC43 has been demonstrated by three independent methods to be present in brain tissue of a case of acute encephalitis.

The HCoV-OC43 strain we identified in the study is similar to a novel OC43 genotype E recently described for the first time (Zhang *et al.*, 2014). In this study of respiratory tract infections, 65 clinical samples were analyzed. The novel genotype is hypothesized to have emerged in 2010 and was identified in five children, all aged less than three years old.

5.3.7 Comparing metaMix results to other methods

We also analysed the dataset using both Pathoscope and MEGAN to compare the performance of the three tools in a clinical sample where there are species absent from the database.

Pathoscope

Pathoscope identified 177 species in this sample. We optimized the value of the unique read penalty parameter and we achieved the best results with the thetaPrior parameter set within the range 10-100. With these settings, the method identified 52 species, including five different coronavirus species (Appendix Table B.4). Our assessment is that Pathoscope is confused by the lack of completeness of databases combined with the absence of an "unknown" category, which prevents it from dealing with these unassigned reads sensibly. Pathoscope completed its analysis in 10 minutes.

MEGAN

MEGAN assigned the reads to 30 taxa. These included some species and genera but most were families. Approximately 250K reads could not be assigned to any taxonomic level. MEGAN ran in 8 minutes.

5.3.8 Concluding remarks

We discussed in chapter 1 the intrinsic limitiation of similarity based methods due to their reliance on the content and completeness of public reference databases. With this case it is emphasized how the database choice impacts the results: the RefSeq database we used has only one HCoV-OC43 strain, while in GenBank there are several, capturing the high mutation rates of this species. Since it is not computationally efficient to use all publicly availble sequences, it is necessary to follow up any similarity based community profiling with further analyses, such as phylogenetic ones for greater resolution and accuracy.

An interesting related point is that we followed up on the sequences assigned to the “unknown category”, looking for nucleotide similarity with NR-NT using BLASTn. Half of the reads originated from an untranslated region of the Coronavirus genome, which is not captured by the protein reference database. The remaining reads matched confidently to either zebrafish or chicken sequences, two organisms whose proteins are not in the custom human microbiome reference we are using. These matches were explained as barcode leakage that resulted from multiplexing on the same flowcell zebrafish and chicken RNA-Seq libraries. metaMix appropriately assigned all these reads to the “unknown” category, producing a clean probabilistic summary. Pathoscope on the other hand, does not have a formal way to handle the absence of closely related sequences in the database and misassigns these reads to species that share very low levels of similarity, resulting in the great number of false positives.

Chapter 6

Viral trigger for Type I Diabetes

In this chapter we discuss the project that was the original motivation for deciding to work on the community profiling research question and for developing metaMix. The exact aetiology for Type I Diabetes (T1D) onset has not been determined, for all the dedicated research effort over the last 100 years. That said, researchers agree that T1D onset results from the interaction between an individual's genetic predisposition, their immune system and various environmental factors (Atkinson, 2014). T1D incidence is increasing in many countries (Harjutsalo, 2008) and it is now believed that one or more environmental factors are driving this increase. Part of this mosaic picture are viruses, especially enteroviruses, whose role in T1D onset has long been implicated (Yeung *et al.*, 2011; Oikarinen *et al.*, 2011, 2014; Richardson *et al.*, 2014a), however the evidence is still inconclusive.

A large collaboration between several institutions across different countries is the JDRF nPOD-Virus group which aims to investigate the role of viruses in T1D. The nPOD-V group brings together researchers with a common interest in viruses and T1D, of diverse expertise from multiple disciplines. The goal is to facilitate the application of multiple methods for enteroviral detection using the same sample set. The nPOD collection (Campbell-Thompson *et al.*, 2012) (`http://www.jdrfnpod.org/`) consists of post-mortem pancreatic samples from organ donors with T1D or at varying levels of risk for the disease. nPOD-V is organised in 6 subgroups (tasks) each with a specific focus. The overall goal of the task we participate in is to apply an integrated approach for RNA analysis towards the identification of viruses associated with T1D. More specifically for our group the aim is to attempt to identify and characterize viruses associated with T1D using high throughput sequencing. nPOD-V investigators from other institutions use alternative methods, such as laser-capture microscopy, real-time PCR and in situ hybridization.

We begin with a section introducing the highlights of the relevant research on the role of viruses in T1D onset. We then discuss the results of the RNA-seq work at each stage of our participation in the nPOD-V project. The sample selection as well as the technology employed throughout the years was informed from the findings in previous stages or other tasks.

6.1 Background

6.1.1 T1D

T1D is characterised by a significant shortage or complete lack of insulin secretion. Insulin is normally produced by endocrine cells which are found in the pancreas. These cells are islands of endocrine tissue distributed throughout the pancreas. One of the cell types that form these islets are the β-cells, which are the ones producing insulin. T1D results from autoimmune destruction of these cells (Bluestone *et al.*, 2010). The resulting insulin deficiency requires daily insulin injections for survival. By the time the diagnosis is made, the β-cells have almost completely been destroyed, making prediction and prevention a high priority (Polychronakos and Li, 2011). Both require knowledge on the causal factors and pathways to disease.

Much of the T1D risk is accounted for by genetic predisposition, with the sibling relative risk (λ_S) estimated to be close to ten (Clayton, 2009). The predominant genetic contribution in humans comes from the HLA complex on the short arm of chromosome 6 (Thorsby, 1997). However, the contribution of other non-genetic factors to the aetiology of T1D is supported by the existence of T1D cases who have developed the disease despite protective genetic loci (Christen and von Herrath, 2011). Furthermore a predominance of susceptibility genes in individuals does not necessarily result in development of T1D (Ziegler and Nepom, 2010). Finally, there is no strict concordance between homozygous twins (Redondo *et al.*, 1999). Also epidemiological evidence shows an annual increase of 3% (Tuomilehto *et al.*, 1999) in T1D incidence in many countries over the past decades (Gale, 2002). This indicates that environmental factors are necessary to initiate and propagate the disease (Bach, 2002).

6.1.2 Viral trigger hypothesis

There is a long standing hypothesis that viral infection can act as a trigger for T1D in genetically susceptible individuals. The most robustly documented relationship between a virus and T1D has been with enteroviruses. Enterovirus is a single-stranded RNA virus and it belongs to the picornaviruses (Coppieters *et al.*, 2012). However there has not been a definite proof for the viral connection. On the contrary the path towards understanding the possible role of viruses in the disease development has been paved with challenges, not least due to the limited availability of pancreatic tissue. Therefore study of even small numbers of pancreatic samples from T1D patients could provide the most convincing argument for the viral link. The proximity of the pancreas to other vital organs has deterred sample collection from living patients newly diagnosed with T1D (Krogvold *et al.*, 2014). Most of the studies described below have been conducted using post-mortem pancreases with one notable exception where pancreatic biopsy was performed on living patients.

6.1.3 Supporting evidence

Extensive literature spanning the past forty years has resulted from the effort to study the role of persistent enteroviral infection of pancreatic β-cells in the initiation and progression of T1D. One of the first noteworthy findings occurred in the 1970s (Yoon *et al.*, 1979) when a coxsackievirus which is a human enterovirus, was detected in the pancreas of a child who died of diabetic ketoacidosis within one week of onset. The virus was injected into mice resulting in islet inflammation, β-cell necrosis and diabetes, with viral antigens detected in the β-cells. Thirty years later the same virus was identified in autopsied pancreatic tissue in half of six T1D patients but in none of the twenty six control organ donors (Dotta *et al.*, 2007). In a larger study in 2009 the enteroviral capsid protein VP1 was detected in multiple islets in the pancreas of patients with T1D who had died within a year of developing the disease (44 out of the 72 patient samples, compared with 3 out of 50 control samples) (Richardson *et al.*, 2009). More recently and using samples from the nPOD collection which consists of more recently harvested pancreatic samples, the same team of authors (Richardson *et al.*, 2013) detected VP1 in 8 of 10 T1D cases that had insulin-containing islets. VP1 was not detected in any of the cases that were deficient of insulin. A conclusion of this study was that enteroviral infection in T1D persists over long periods rather developing as an acute infection. Enterovirus infections usually proceed very rapidly (Richardson *et al.*, 2011) and this unusual pattern of only a few islet cells in T1D becoming infected with this low-level mode may be important in triggering islet autoimmunity. The conclusion was reached on the basis that the nPOD cases were from T1D patients where the disease duration was longer, ranging from one to twenty years and mean duration of twelve years.

It must be noted that the antibody widely employed to detect the enteroviral VP1 in islet cells in immunostaining studies might also cross-react with additional proteins under some conditions (Richardson *et al.*, 2013). It is therefore crucial that further evidence is acquired before deducing that the immunodetection of VP1 in T1D case islets is associated with an underlying viral infection.

A different type of supporting evidence comes from a genome-wide association study (GWAS). Specifically it has been found that there are four rare variants of *IFIH1* (*interferon-induced helicase 1*) that independently decrease the risk of T1D through a lost or reduced expression of the protein (Nejentsev *et al.*, 2009). Additionally disabling *IFIH1* expression lowers the risk of T1D (Nejentsev *et al.*, 2009). *IFIH1* is an interferon response gene which allows the infected cell to sense RNA viruses and increase interferon production by the host immune system. Interferons are signalling proteins that trigger the immune system's defence in order to limit viral replication and prevent damage to the infected cell. However they also increase the visibility of the infected cell to the immune system making it highly susceptible to recognition and destruction by cytotoxic CD8 T cells (CD8+ T cells). CD8+ T cells recognise an antigen when it is presented to them bound to cellular class I Major Histocompatibility Complex (MHC) molecules. This presentation is enhanced when there is heightened expression of class I MHC, commonly seen in the islets of patients with type 1 diabetes (Foulis *et al.*, 1987). This has been hypothesized to be one mechanism by which β-cells become visible to the immune system during the development of autoimmunity and can be summarised as the "fertile field hypothesis" (von Herrath *et al.*, 2003), whereby the virus infects the β-cells and predisposes them to autoimmune attack (Green *et al.*, 2004).

It is worth noting that recently researchers participating in the Diabetes Virus Detection (DiViD) project used pancreatic biopsies of six living individuals with recent

T1D onset (3-9 weeks) (Krogvold *et al.*, 2015). Enterovirus was detected in all cases by either RT-PCR (four positive for viral RNA out of six cases) in two different laboratories or IHC (islets cells in all cases positive for viral protein VP1). The amount of enterovirus RNA was low but its presence was additionally confirmed by sequencing the PCR products. RNA-seq data for each patient were generated using an Illumina HiSeq 2000 instrument. No viral reads were identified using RINS (Bhaduri *et al.*, 2012), an approach where reads are mapped to a pathogen database, matching reads are assembled into contigs and finally all original reads are mapped on these contigs. The study was ceased by the DiViD investigators as there were complications for three of the patients, who ultimately recovered fully. While the biopsy procedure was uncomplicated for three participants, the complications that arose for the remaining volunteers included extensive post-operative bleeding, pancreatic drainage, splenic tear, pain and fever (Krogvold *et al.*, 2014), resulting in extended hospitalisation.

It is clear that the safest and easiest means for obtaining access to pancreatic tissue is by using post-mortem samples. These are also rare but our participation to the nPOD-V group has allowed us to get access to 30 nPOD cases, approximately half of them T1D cases over the course of 4 years. In the following section we discuss the results obtained by the analysis of the RNA-seq data we generated from the nPOD samples. The sample selection at each stage was guided by the results emerging from other tasks as well as previous stages.

6.2 Results

Samples from cases with short and longer disease duration were selected for sequencing. Additionally preclinical T1D cases, that is cases without the disease but in high risk for developing the disease were also included in the selection. The high risk is indicated by the presence of multiple antibodies produced by the immune system that attack the body's own cells, tissues and organs causing inflammation and damage, called autoantibodies. Presence of persistently positive and multiple autoantibodies is highly predictive for the development of T1D (Pihoker *et al.*, 2005).

The criteria for sample selection from the nPOD collection evolved through the years and as the collection increased in size. During Stage I and Stage II these were simply availability of cases while during the later stages (III and IV) T1D cases were selected on the basis of VP-1 immunohistochemistry (IHC) positivity, i.e they were positively stained for viral antibodies. Additionally hyper-expression of class I MHC, which as discussed above is common in the islets of T1D cases, was taken into account for Stage III and Stage IV.

In all generated data so far the results are negative for the presence of enteroviruses. However this negative result should be interpreted with caution as appropriate sample collection is challenging. Common problems include the very low number of infected islets in a T1D pancreas, the timing of collection, the fixation with formalin the post-mortem tissue undergoes which impacts greatly the RNA stability and integrity (Richardson *et al.*, 2014b).

The bioinformatics pipeline and metaMix evolved over the years along with this project, however all results presented here have been reanalysed with the latest pipeline verion (January 2015) and metaMix 0.1.

6.2.1 Stage I - Prior to nPOD-V

Prior to the formation of the nPOD-V group, we worked on two preclinical T1D cases which means they were autoantibody positive, from the nPOD collection. Three deep sequencing datasets were generated from the two samples using two different library preparation techniques, poly(A)-purification and ribodepletion, discussed in Chapter 1.

Acquiring samples of high quality can be challenging, especially for post mortem clinical samples, despite standard RNA sample handling procedures (Sigurgeirsson *et al.*, 2014). Due to the nature of these samples, the quality of RNA in terms of degradation was generally low. This is indicated by a measure called RIN (RNA Integrity Number) (Schroeder *et al.*, 2006), produced by an algorithm that can detect presence of degradation products. The RIN score ranges from 1 to 10, where level 10 denotes completely intact RNA. Therefore the lower the score, the lower the quality of the RNA (Schroeder *et al.*, 2006). In general, RIN values greater than 7 are considered good quality.

Unfortunately for these samples we were not able to recover the RIN scores. However judging from the extensive degradation of all follow up samples, we can assume these would be low as well. RNA-sequencing was carried out using Illumina GAIIx. The reads count summary are in Table 6.1.

Table 6.1: Stage I prenPOD-Virus - 2 nPOD samples, 3 datasets: deep sequencing reads summary statistics. The asterisk denotes the dataset produced by the ribodepletion approach.

CaseID	**Donor**	**Raw Data (in K)**	**Unique %**	**QC %**	**Non-Host %**	**Non-rRNA %**	**Protein simil. %**
6044	preclinical T1D	34538.3	17.5	17.1	1.1	1.1	0.4
6044*	preclinical T1D	24014.2	18.8	18.3	1.0	1.0	0.3
6027	preclinical T1D	67391.4	12.1	11.9	0.8	0.8	0.3

The community profile for the three datasets consisted of 132, 95 and 157 taxa respectively. The majority of the reads remained unassigned (a proportion greater than 70% in all samples) with the remaining reads assigned to several bacteria, each supported by a low number of reads. These are either real bacteria that exist in the pancreatic tissue or they are contaminants originating from PCR reagents and/or extraction kits (Salter *et al.*, 2014) or human-skin associated bacteria acquired during sample handling. Pancreas is however thought to be a sterile tissue (Funchain and Charis, 2012) and therefore contamination appears to be the more plausible explanation, even though negative controls would be necessary to confirm or refute this.

The number of taxa is not necessarily surprising, as the pancreatic samples contain a low microbial biomass, allowing the contaminating sequences from the extraction kit or the reagents or the general lab environment to overtake a larger fraction of the sequences. Additionally due to the low read numbers originating from each contaminant, there is not enough information for metaMix to definitively differentiate between different strains of the same species. In such a situation metaMix retains all of the competing strains, increasing the number of reported taxa in the profile. Increasing the read support parameter to a larger value would result in more parsimonious summaries. As an example, we ran metaMix again using `r`=50 and this reduced the number of identified taxa to 57, 37 and 61 respectively. However our analysis requires a low detection limit as the goal is to detect traces of viruses that may be potentially present, therefore we performed all metaMix analyses using the default value of `r`=10.

The two datasets from case 6044 showed an important difference. In the ribodepleted dataset a murine leukemia virus (MLV) was identified. This was due to contamination of laboratory reagents with mouse retroviruses, a known issue (Robinson *et al.*, 2010; Oakes *et al.*, 2010) which was discussed in chapter 1 . This result is also briefly discussed in Appendix A (Section A.2, Figure A.4) as the discovery that helped me realise the importance of utilising the available information from all reads in contrast to considering only the assembled contigs.

6.2.2 Stage II

During the first official stage of the nPOD-V group in spring 2012 we sequenced seven post-mortem pancreatic samples from four T1D patients, one T2D and two healthy subjects as negative controls. The duration of the disease ranged from 4 to 28 years.

Three of the pancreatic samples had a RIN score of 2-3, so these were pooled to be sequenced together. The remaining four samples were moderately degraded with a RIN score above 4 and only one sample having a value greater than 7. Despite this we proceeded with RNA-sequencing using Illumina GAIIx. The same sample set was also sequenced at Baylor College of Medicine (BCM), where the library preparation step included human ribosomal RNA (rRNA) depletion.

I used the bioinformatics pipeline described in chapter 1. Read statistics are summarised in Table 6.2 and plotted in figure 6.1. Even though the number of raw reads is high, there is a great degree of clonality due to PCR amplification, suggesting that the complexity of the library had been exhausted. The informative reads for species identification which are the ones sharing some similarity to the proteins in the reference database, are a few hundred thousands for all 5 datasets, ranging between 207,000 and 377,000 reads.

Table 6.2: Stage II - 7 nPOD samples: deep sequencing reads summary statistics

CaseID	**Donor**	**Raw Data (in K)**	**Unique %**	**QC %**	**Non-Host %**	**Non-rRNA %**	**Protein simil. %**
6070	T1D	65134	18.4	17.9	0.8	0.7	0.6
6098	Control	68481	17.7	17.3	0.6	0.6	0.5
6127	T2D	64459	17.8	17.4	0.6	0.6	0.5
6141	T1D	67146	16.9	16.5	0.6	0.6	0.4
6046,6084,6099	T1D, T1D, Cntr	72379	31.9	30.7	0.5	0.5	0.3

We subsequently applied metaMix to the annotated reads and contigs. In general for all samples excluding the pooled dataset, a very similar profile was produced. No viruses of interest were found - the majority of reads originated from Enterobacteria phage phiX174, the positive control for Illumina sequencing, while the rest of the reads were divided between various contaminants and the “unknown” bin (Table 6.3). The

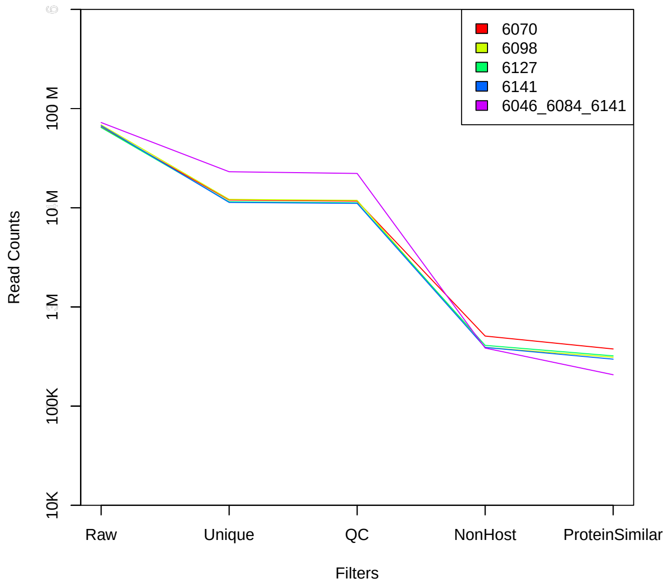


Figure 6.1: Stage II nPOD samples: number of reads passing each filter stage. The reads that show some protein similarity are in the low hundred thousands range.

profile was different for the pooled dataset where half of the sequences could not be assigned confidently to any of the species in the reference database, while the other half originated from bacterial contaminants. A representative metaMix summary for these samples can be found in Appendix Table B.6.

Table 6.3: Stage II - 7 nPOD samples: General profile and relative abundances in 6070, 6098, 6127, 6141.

organisms	mean abundance % (sd)
Enterobacteria phage phiX174	62 (0.03)
Environmental bacteria	25 (0.02)
unknown	13 (0.01)

The community profile in all cases consisted of 80-90 taxa. Similar as before the reagent contaminating sequences dominate numerically the results and metaMix returned several strains of the same bacterial species. As an example it was common to find different substrains of the *Escherichia coli* species, most at very low levels (less than 100 reads). E. coli is a known contaminant of PCR reagents (Silkie *et al.*, 2008). However the same E. coli signal was detected in the resulting datasets from BCM, indicating that the contamination had occurred before the samples were dispatched to the two different sequencing centers. The more plausible explanation is that rather than different E. coli strains being simultaneously present in the samples, metaMix cannot distinguish between them using the specified read support value of 10 reads. Subsequently all the strains that have at least ten unique reads assigned to them - or equivalently a greater number of ambiguous reads that in sum provide as much information as ten unique ones - are retained in the present species summary. The poor read classification probabilities for most of the E. coli strains further supports this interpretation, underlining the lack of clear signal for all of them. The classification probabilities for two of the E. coli strains in sample 6127 are plotted below in Figure 6.2. For *Escherichia coli SE11*, 6.5K reads have a classification probability greater than 0.8 In contrast, for *Escherichia coli O157:H7 str. EDL933* only 25% out of the 2K reads do so, while the remaining 1500 reads have lower probabilities, indicating they match better or equally well other bacterial species, presumably the other E. coli strains.

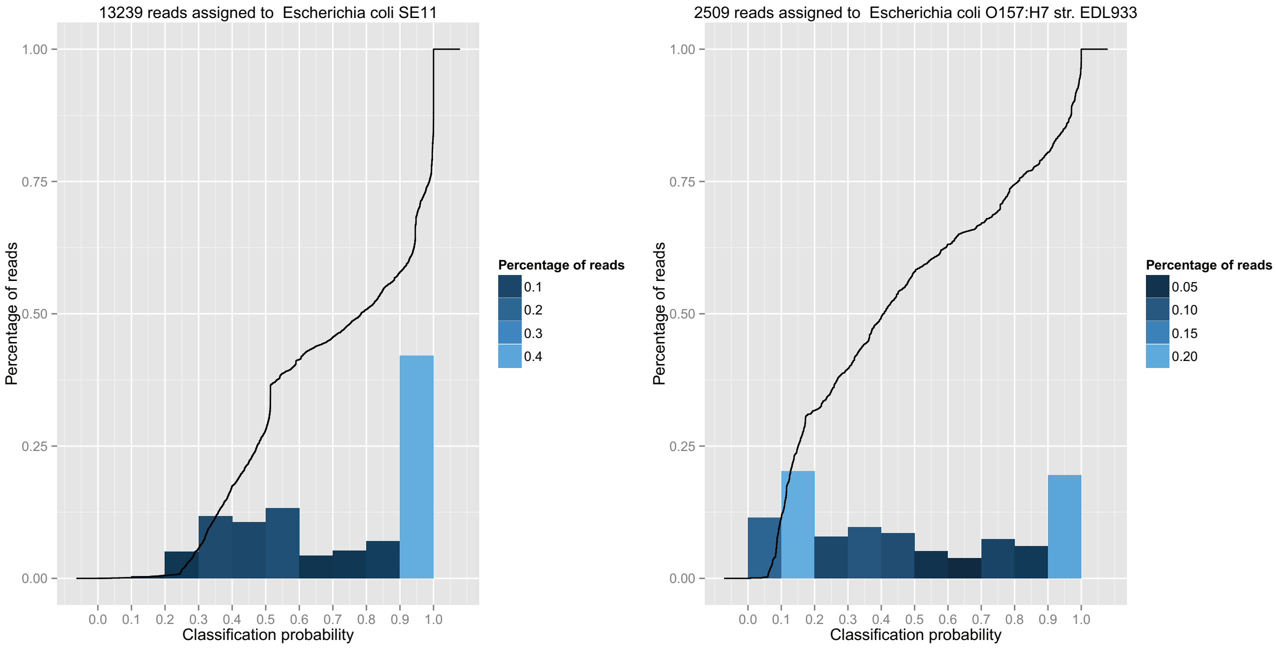


Figure 6.2: Stage II - Classification probabilities for two out of ten *Escherichia coli* strains in the results of sample 6098.

A second round of annotation for the "unknown" reads followed, using BLASTn and the NR-NT database. This is a nucleotide sequence database with a significantly greater number of sequences, as it contains entries from both RefSeq, EMBL and GenBank. Approximately 15%-25% of the unknown reads in all datasets showed good similarity to an E. coli strain, while the rest did not have any matches in the extended database. An alternative approach to consider for future use would be to try to characterize the unknown reads with PSI-BLAST (Position Specific Iterative BLAST) (Altschul *et al.*, 1997) which is run in multiple iterations. The first iteration uses the results of a BLAST search to create a position specific score matrix. This matrix is used in the subsequent iterations, instead of the standard scoring matrices for protein similarities, to generate more specific results. A different path could be explored by trying to extend by targeted PCR and further sequencing the unclassified contigs in the unknown category. This may then allow the detection of distant homology to a known virus.

6.2.3 Stage III

The next stage of the project involved deep sequencing of 12 pancreatic slices. The goal was to employ again the direct approach used in Stage II, while increasing the sample size with cases of a shorter disease duration. The shorter duration implies that samples were collected closer to the onset of the disease and in turn closer to a potential triggering viral infection. There were four T1D cases, with disease duration ranging from 1 year to 4 years. The remaining 8 samples were from cases with high VP1 immunohistochemistry positivity.

The samples were degraded with RIN values ranging from 2.5 to 6.7. Same as before due to the uniqueness of the samples, we proceeded with RNA-sequencing switching to Illumina HiSeq2500. Read statistics are summarised in Table 6.4 and plotted in figure 6.3. Similar to Stage II, the proportion of independent (non clonal) reads is low, while the reads showing homology to proteins range between 30K and 700K.

Table 6.4: Stage III - 12 nPOD samples: deep sequencing reads summary statistics

CaseID	Donor	Raw Data (in K)	Unique %	QC %	Non-Host %	Non-rRNA %	Protein simil. %
6052	T1D	184527	38.8	21.3	0.8	0.5	0.3
6080	Autoab Pos	167442	38.2	23.8	0.8	0.6	0.4
6090	Autoab Pos	107271	13.5	13.2	0.1	0.08	0.05
6113	T1D	83012.8	18.1	16.4	1.7	1.5	0.8
6147	Autoab Pos	99935.6	12.6	12.3	0.06	0.05	0.03
6158	Autoab Pos	89958.2	15.5	14.6	0.2	0.2	0.1
6167	Autoab Pos	154587	39.1	25.4	1.14	0.7	0.4
6171	Autoab Pos	96542.8	17.9	16.3	1.7	1.5	0.8
6195	T1D	101691	12.0	11.9	0.06	0.05	0.03
6197	Autoab Pos	85471.3	18.3	16.9	1.3	1.2	0.6
6198	T1D	89723.8	16.7	15.2	1.1	1.0	0.5

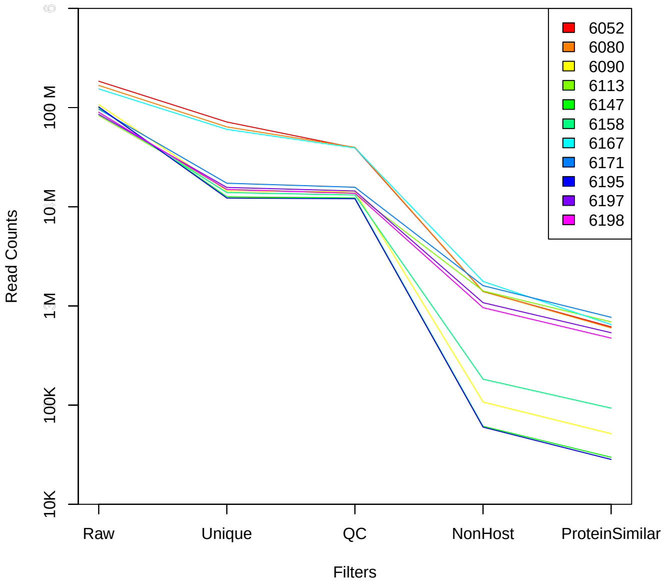


Figure 6.3: Stage III -12 nPOD samples: number of reads passing each filter stage. The reads that show some protein similarity are in the low hundred thousands range.

In all 12 cases the majority of the reads (80-90%) could not be assigned by metaMix to any species. Further analyses on the unknown reads were performed same as in Stage II, including BLASTn search of the NR-NT. The vast majority of the reads in the “unknown” bin remained unclassified while a small minority had matches to bacteria, either human skin associated or associated with contamination of laboratory reagents and extraction kits. The remaining reads were assigned to either *Enterobacteria phage phiX174* (5-10%) or environmental bacteria (less than 1%). In three samples (6052, 6080, 6171) the *Geobacillus virus E2* was detected at low levels. Given its homology to the Bacillus species which is a known contaminant (Salter *et al.*, 2014), we consider it as not relevant to the viral trigger pursuit.

In general for all samples the metaMix profiles consisted of 40-60 taxa, with the similar pattern of several strains of the same species appearing in the results. The metaMix summary for case 6198 can be found in the Appendix Table B.7 and is representative of the profiles in all cases.

6.2.4 Stage IV

The failure to detect viral nucleic acid using direct sequencing of the low volume pancreatic samples in Stage II and Stage III prompted us to consider a different method. A more suitable approach is a target capture method for purifying low quantities of viral nucleic acid from samples where the host genome forms the majority. This technique is used extensively in the Breuer lab. Therefore for the final stage of our involvement with the nPOD-V project, we collaborated with Judy Breuer and Daniel Depledge who developed the method for isolating and enriching for specific viral genomes prior to deep sequencing (Depledge *et al.*, 2011). The sequence capture method enriched the RNA samples for enteroviral genomes which are the target sequences of interest.

6.2.4.1 Stage IVa

We first assessed the sensitivity of the method on five samples that were spiked-in with *Coxsackie B virus* (CBV1) at different dilutions and one negative control. The samples were multiplexed and run on a Illumina MiSeq. The results from this experiment indicated a linear relationship between the dilution level of the spiked-in virus and the number of CBV1 reads detected. The read summary is in Table 6.5.

The presence of one pair of CBV1 reads in the negative control suggests low level cross-contamination occurred between samples with high viral load. This was a bit worrisome as we generally are interested in detecting low numbers of reads; that said one pair of reads would not in general be considered as enough evidence to believe in the virus presence. Furthermore, one pair of reads would not provide sufficient support for metaMix to keep CBV1 in the present species. Finally, we expect low amounts of the virus in the actual pancreatic samples and therefore carry-over between samples would be unlikely.

Table 6.5: Stage IVa - Control experiment: five spiked-in samples with CBV1 and one negative control

CBV1 **spike-in**	**Raw** **# pairs**	**QC** **# pairs**	**QC** **%**	**CBV1 aligned** **# reads**	**CBV1** **%**
negative control	4728119	4349094	0.92	2reads→1 pair	2.3E-007
10e-8	4223286	3971535	0.94	32	4.0E-006
10e-7	3710861	3351818	0.90	283	4.2E-005
10e-6	3546996	3229128	0.91	2595	4.0E-004
10e-5	4392146	4015026	0.91	12237	1.5E-003
10e-4	2448108	2149775	0.88	132770	3.1E-002

6.2.4.2 Stage IVb

Following the promising results from the positive control experiment, the next step was to apply the enrichment method on further 12 nPOD samples. Of these six were T1D cases and six were post-mortem pancreatic samples from healthy subjects (controls). The T1D cases were selected on the basis of positive VP1 staining via IHC and MHC class I overexpression. The RIN scores revealed extensive degradation with values ranging from 2.1 to 3.5 and only one sample (6213) having a score of 6.5. The samples were sequenced on an Illumina NextSeq machine. Read statistics are summarised in Table 6.6 and plotted in figure 6.4. Same as in Stage II & III, we observe high clonality in the resulting datasets.

Table 6.6: Stage IVb - 12 nPOD samples: sequence capture technology - deep sequencing reads summary statistics

CaseID	**Donor**	**Raw Data (in K)**	**Unique %**	**QC %**	**Non-Host %**	**Non-rRNA %**	**Protein simil. %**
6024	Control	63911.1	21.2	21.1	1.6	1.4	0.4
6052	T1D	73753.2	18.5	18.4	1.2	1.1	0.3
6055	Control	48735.2	21.3	21.2	1.4	1.3	0.4
6073	Control	73871.1	19.6	19.5	1.4	1.2	0.4
6095	Control	67890.3	19.3	19.2	1.3	1.1	0.3
6126	Control	65789.9	20.6	20.5	1.5	1.4	0.4
6165	Control	68196.9	19.6	19.5	1.4	1.2	0.4
6213	T1D	70283.4	19.0	18.9	1.4	1.2	0.4
6228	T1D	72890.7	19.5	19.3	1.3	1.2	0.4
6243	T1D	72498.4	19.7	19.6	1.3	1.1	0.4
6070_02	T1D	76688.9	18.6	18.5	1.4	1.2	0.8
6070_04	T1D	70321.8	19.1	19.0	1.4	1.2	0.4

Community profiling with metaMix identified only two species in the results, with the vast majority of reads - greater than 95% in all samples - being unassigned. One or two thousand reads were host sequences while less than a thousand reads assigned to *Enterobacteria phage phiX*. A typical summary can be seen in Table 6.7.

Table 6.7: metaMix summary profile for case 6055, deep-sequenced using the sequence capture approach

taxon id	**scientific name**	**assigned reads**	**posterior prob**
unknown	unknown	188927	1
9606	*Homo sapiens*	2745	1
374840	*Enterobacteria phage phiX174 sensu lato*	698	0.92

We considered the possibility that unassigned reads could actually originate from

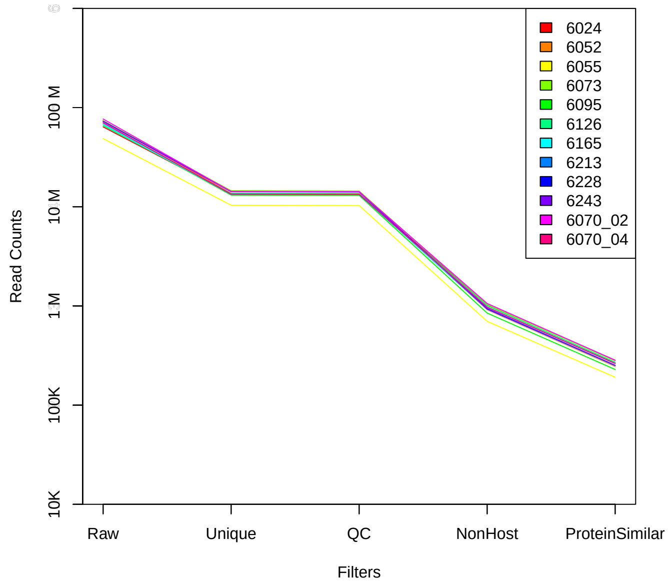


Figure 6.4: Stage IVb - 12 nPOD samples: number of reads passing each filter stage.

a species that was not present in our reference database. We subsequently performed a BLASTn search against the nucleotide-NR database but the reads remained unclassified, with only a few thousand reads matching human, bacterial or uncultured eukaryote sequences.

In two of the samples we detected the presence of Human Herpesvirus 3 (HHV3), supported by 18 and 69 reads respectively. The classification probabilities indicated this was a true signal and that the sequences were indeed originating from HHV3 (Figure 6.5). This is a pathogen of interest in the Breuer lab, frequently sequenced on the NextSeq and MiSeq machines we used for data generation during Stage IV. Thus the most likely explanation would be cross-contamination between runs or during library preparation steps.

6.2.4.3 Stage IVc

Over the years and thanks to relevant research the importance of sequencing the right cells became obvious. This guided the decision to generate RNA-seq data from laser microdissected nPOD donor islets. An initial run on the MiSeq has been performed for these samples. It is also planned to sequence these in greater depth using the NextSeq instrument. The data summary is in Table 6.8 and plotted in Figure 6.6 and we see that the reads showing similarity to proteins are only a few thousands.

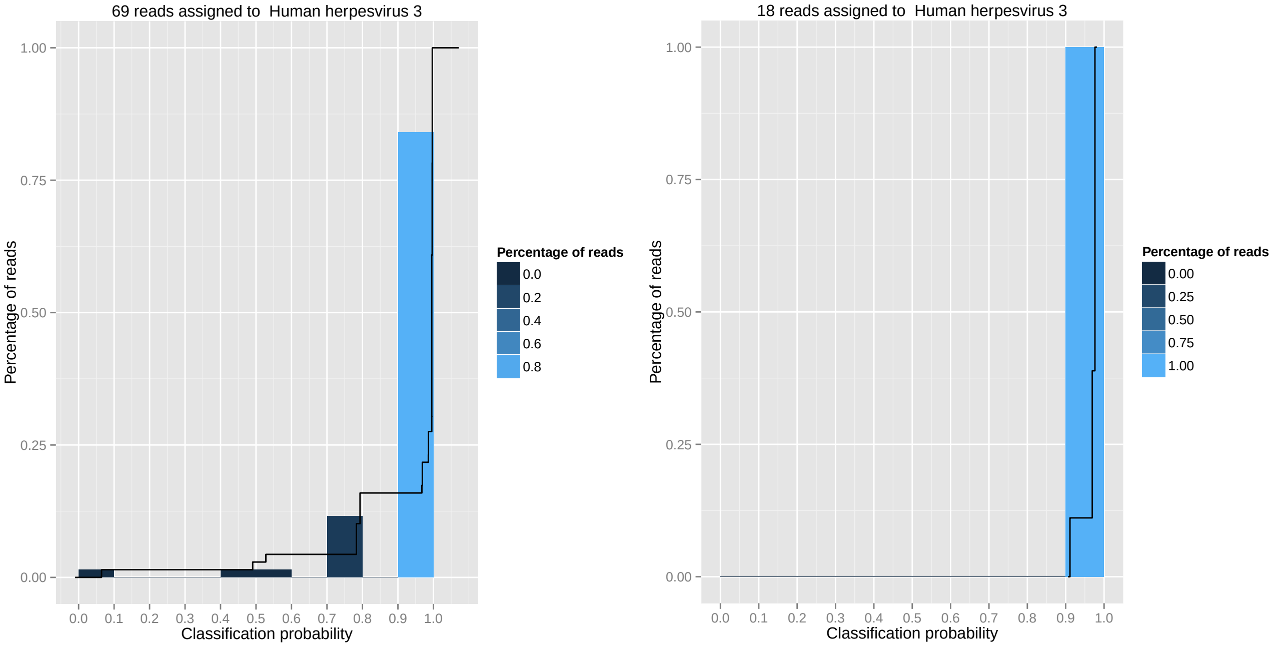


Figure 6.5: Stage IVb - Classification probabilities for detected HHV3 in samples 6052 and 6165

Table 6.8: Stage IVc - 4 laser capture microdissected islets samples: sequence capture technology - deep sequencing reads summary statistics

CaseID	Donor	Raw Data	Unique	QC	Non-Host	Non-rRNA	Protein
		(in K)	%	%	%	%	simil. %
6052	T1D	1523.89	31.01	30.27	0.48	0.39	0.21
6070	T1D	1942.9	20.51	19.97	0.41	0.35	0.21
6213	T1D	1152.03	24.93	24.23	0.59	0.48	0.27
6243	T1D	1078.69	34.47	33.24	0.56	0.45	0.24

The metaMix profile consisted only of two, three species per sample. The familiar outcome was that the vast majority of the reads remained unassigned, however a few hundred reads were assigned to Human Herpesvirus 1 and 3, in three out of four samples. The metaMix summary for case 6213 (Table 6.9) has been provided as an example.

Table 6.9: metaMix summary profile for laser dissected islet from case 6213, deep-sequenced using the sequence capture approach

taxon id	scientific name	assigned reads	posterior prob
unknown	unknown	3726	1
10298	*Human herpesvirus 1*	227	1
10090	*Mus musculus*	87	1
10335	*Human herpesvirus 3*	4	1

Both these pathogens are extensively sequenced in the Breuer lab where our se-

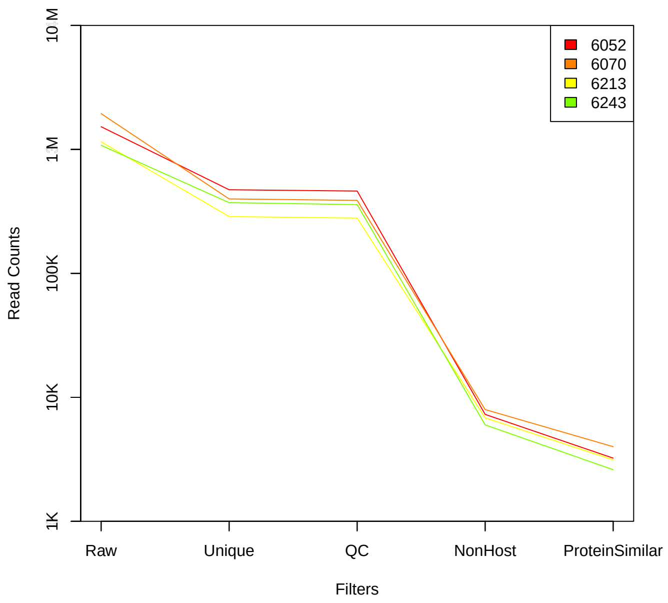


Figure 6.6: Stage IVc - 4 laser capture microdissected islets samples: number of reads passing each filter

quencing happened so it was indicative of contamination. This was due to a cross-run contamination problem of the MiSeq where a small residue of the last library is sequenced in the subsequent run. The problem is now well-known and has been resolved by adopting a new wash procedure between runs using bleach, as suggested by Illumina.

6.3 Concluding remarks

Our involvement in this project is coming to an end but additional experiments are planned. In the first instance the plan includes deeper sequencing of the the laser-dissected islets. An additional experiment is to sequence a different set of positive controls from mice where the pancreases have been infected with the virus. This would allow a more realistic simulation that more accurately reflects the biology of islet infection by enteroviruses.

We previously discussed that due to the small true microbial biomass in these post mortem pancreatic samples, the probability that contaminants occupy a big proportion of the dataset sequences is high. This was demonstrated by the typical community profile for these samples, with high numbers of bacterial taxa in the results, each supported by few reads. Other sources of laboratory-specific contamination also plagued some of the data. Ideally we want to conclusively confirm that the present species are contaminants and we could have achieved that with the use of blank negative controls. The lack of incorporated negative controls renders a smooth interpretation of the deduced profile more difficult, but the crux of the matter is not affected. In the end and for all the progress that has been made since the start of the nPOD collaboration, the main question of interest, that is whether there is a viral trigger for the onset of T1D, is still unanswered. The rate of scientific discovery is delayed by the limited or complete lack of access to appropriate specimens for study.

Chapter 7

General Conclusions

In this final chapter, I restate the research objectives of this thesis and outline the novel methodological ideas and their development. I also discuss the thesis research achievements, limitations and possible future work directions. I also briefly outline my thoughts on metagenomic deep sequencing in the service of medical diagnostics as well as its strengths and limitations.

7.1 Thesis objectives: achievements, limitations and future work

Community profiling of a metagenomic mixture can be defined as the identification and quantification of the present species in a sample. A profile is usually obtained by assigning the sequencing reads to different taxa. This is a challenging problem which raises complex computational issues. Similarity based methods use algorithms such as BLAST and are considered the most accurate methods for read assignment and classification. The relatively short read lengths of existing deep sequencing technologies and the homology across viral and bacterial species is part of the reason the current state-of-the-art methods struggle with ambiguous matches and false positives supported by few sequencing reads each. This common problem may cause users to ignore the low abundance organisms in the resulting summaries as these may be false positives. Depending on the research context, this can be a logical approach to bypass false positives. It can also be troublesome in situations where the detection of a potentially disease causing pathogen depends on identifying traces (very low number of reads) in the dataset. The pathogen contribution to the mixture depends on the biological context, the timing of sample extraction and the type of pathogen considered.

Accurate viral identification and quantification in an otherwise complex mixture where other eukaryotic and prokaryotic sequences exist, has been relatively unattended despite the wealth of profiling and binning methods. Most of the similarity based approaches use nucleotide similarities for inference which is suboptimal for virus detection as discussed previously, or have only recently added support for viral marker genes. A limited number of the methods provide support for BLASTx-type searches. Even these however may exhibit low specificity due to the fact that the comparative plausibility of different profiles is not considered. This becomes especially pronounced when discovery of low abundance viral pathogens is of interest, when species are absent from the database, or when some closely related strains exist in the sample, as shown in Chapters 4 and 5. Therefore, a sensitive and specific computational approach was required in such settings for addressing the aforementioned concerns in a formal and elegant way.

This thesis has presented metaMix, a novel method that unified two ideas: the use of mixture models for classification of a single read by borrowing information from the whole of a dataset and the use of parallel MCMC for exploring the state-space of different profiles, by comparing different combinations of species based on their posterior probabilities. This is a computationally intense task as even with a relatively small number of species to consider, the number of subsets that could explain the mixture grows exponentially. Our strategy is based on Parallel Tempering, a Monte Carlo Markov Chain technique, using parallel computing to speed up the inference. An important feature of the method is that it provides probabilities that answer pertinent biological questions, in particular the posterior probability for the presence of a species in the mixture as well as the relative evidence in the data for the presence or the absence of species, as captured by Bayes Factors. A consequence of the increased accuracy is that metaMix produces better estimates for the relative abundances of the organisms in the mixture.

The working assumption underlying the development, testing and fine tuning of metaMix was that we are interested in detecting minute amounts of viruses in the samples, usually from sterile human tissue. This also the reason why the input data for metaMix are derived from an amino acid similarity search step. Such samples are better described by parsimonious community profiles. The profile may be augmented by microbial contamination which is possible at various stages of sample handling and processing and it is important that reads originating from contaminants are not misclassified. The method can deal with either unassembled reads or assembled contigs or both, allowing for flexibility of choice for the bioinformatics preprocessing. In practice, the choice of bioinformatics processing prior to the application of the Bayesian mixture analysis must be optimized for each application. The processing pipeline used in this thesis has been designed with viral sequence identification from RNA-sequencing as a main goal.

metaMix has been assessed on clinical datasets from sterile tissue, its performance compared to that of other methods addressing the community profiling task. For the clinical cases the ground truth is naturally not known. However, the presence of the potential viral pathogens was confirmed with further experiments. Additionally, the parsimonious profile produced by metaMix fits logically with our current knowledge on communities in sterile tissues. These datasets proved to be challenging for the other state-of-the-art methods. All exhibited lower specificity than metaMix presenting a larger number of taxa in their profiles which, given the nature of these samples, are most likely false positives. Some of the methods struggled particularly with viruses, introducing a plethora of viral species in their results. Others failed to cope with reads originating from species not in the database, which resulted in misassignments.

On the other hand, a first limitation of metaMix is the *Poisson* probability approach for the estimation of the generative p_{ij} probabilities using the number of mismatches. Consideration of the differences between mismatches and the fact that some substitutions are more likely than others, especially on the amino-acid level, can lead to more accurate estimation of the p_{ij} probabilities. Such information is captured by the BLASTx similarity score, whose informational wealth is not used to its full extent by metaMix. A modification in the probability estimation that uses this metric has other added benefits, such as automatically incorporating information on the length of the read sequence and the target sequence. This will also be useful for contexts where nucleotide, rather than protein search is the optimal strategy. Human clinical samples where the infectious agents of interest are most likely bacteria benefit more from nucleotide comparisons and target sequence length has to be taken into account.

Relevant to this point, even though metaMix has been used extensively with the type of samples described above (30 nPOD samples and several brain biopsies since the end of the PhD), it has not been applied nor tested on different types of data. This includes datasets originating from metagenomics samples open to the environment (skin, throat swabs, stool) where a plethora of bacterial species are expected or where the infecting pathogen is a bacterial species. The only exception is the FAMeS benchmark datasets which consist of real Sanger senquencing reads which are combined to simulate different mock communities (bacteria, plethora of organisms and some related strains, different combinations of dominant populations) to the ones we typically analyze. Albeit different, the FAMeS datasets were selected during the metaMix benchmarking stage as there were very few datasets for which the exact ground truth was known in terms of both composition and abundance that we could use for method assessment and comparison, and are widely known and used for testing. This proved that the same ideas can work for different datasets to the ones metaMix was designed and optimised for. metaMix outperformed the other methods in terms of better balance of sensitivity and specificity. Still, additional work will be required for optimisation of metaMix for different settings to the ones it was created for. For example, the fixed generative probability for the “unknown” category will need to be estimated again for such datasets, in order to account for the significant differences between the genome lengths of human, bacteria and viruses. Ultimately, the metaMix extension will benefit mostly from the more accurate p_{ij} estimation that will utilise the BLAST score. Other modifications will be required in the bioinformatics preprocessing pipeline, such as retaining ribosomal RNA which would now provide useuful information or as mentioned usage of BLASTn instead of BLASTx.

I have strived to make this method as sensitive and specific as possible while maintaining the efficiency that would allow its use on larger scale data sets. Again, this refers to datasets where the majority of the read sequences is removed as human and thus leaves a manageable amount (up to one million reads) for community profiling. The high sensitivity and specificity of metaMix comes at an increased computational cost, requiring access to a multi-core computer to run efficiently. For the datasets presented here, the computation time remained manageable and did not exceed a few hours, using twelve cores to run twelve parallel chains. Nevertheless, the second obvious limitation of metaMix is the increased processing time for very large datasets. Speed related improvements can be implemented in scenarios where the species ambiguity concerns only a small proportion of the read set. Reads with certain assignments can be flagged prior to the MCMC exploration of the state-space. Their assignment information can then be carried forward, thereby reducing the size of the similarity matrix used as input by the mixture model. Another area of possible improvement is MCMC convergence determination. The current version of metaMix produces log-likelihood traceplots allowing the user to visually inspect the MCMC convergence, however additional diagnostic criteria can be implemented in future versions.

A final limitation of metaMix which is universal for similarity based methods, is its reliance on the content and completeness of public reference databases. As mentioned previously, public databases hardly represent the real biological diversity, especially pertinent for the viruses that are mostly undiscovered. Additionally their content is biased towards cultivable organisms and human pathogens. Therefore reads from novel microorganisms that are sufficiently divergent from known species will remain unclassified. The users of metaMix can easily obtain the unclassified reads for follow up investigations regarding their nature. The classifications and resulting profile may differ depending on the choice of the database. Informed database selection will improve results, however users should always bear in mind the inherent limitations, biases and potential errors in order to avoid misinterpretations of unlikely findings.

7.2 Deep sequencing for medical diagnostics

Deep sequencing has been a tranformative technology, affecting the whole breadth of the biomedical sciences. The potential for massive parallelisation and automation has made large scale sequencing projects possible. It has reshaped the field of metagenomics as it provides the opportunity to sequence uncultured microorganisms sampled directly from their natural habitats.

Metagenomics can detect bacteria, viruses, fungi and parasites simultaneously. RNA viruses are detected with the use of metatranscriptomics, however for the remaining of the discussion we will not make this distinction and use the term metagenomics for both approaches. A promising application of metagenomics can be found in medical diagnostics, where high throughput sequencing is starting to revolutionise pathogen detection and to inform treatment strategies. Widely used traditional techniques for pathogen detection can be time consuming - viruses in particular are difficult and often impossible to culture - and require prior information on the potential infectious agents. This hampers the detection of unsuspected or undiscovered pathogens. Metagenomics offers a target-independent approach for pathogen detection and no prior knowledge on the cause of the infection or outbreak is required. Additionally it offers the possibility to characterise both individual organisms as well as the community in the sample. Other properties of a pathogen such as virulence and drug resistance may be uncovered.

Diagnostic virology has already benefitted by the successful use of metagenomics.[1] Quick and target independent viral identification and discovery was previously hindered by the difficulty associated with culturing viruses and their lack of a universally conserved genetic element shared between viral genomes. Challenges such as the emergence of novel strains have been resolved with the use of metagenomics. Accurate identification is crucial for avoiding misdiagnoses that may lead to improper clinical treatment and which negatively affects survival or transmission rates. The detection and response to viral pathogen outbreaks is another application of metagenomics, successfully used for example in influenza outbreaks to determine viral subtype. Another attractive feature of deep sequencing is the ability to detect variants at low frequencies. This is useful for identifying drug resistant mutations or transmission patterns of the viruses and for evaluating the impact of minority variants on treatment efficacy. Finally, a number of viruses that have not been associated with human diseases have been detected in healthy human hosts, establishing the existence of a normal human virome.

There are various challenges and bottlenecks that need to be addressed before routine application in diagnostic capacity is feasible. These include computational limitations that are not only related to data processing but also to data storage. Additionally, access to state-of-the-art computer equipment is not necessarily straightforward for hospitals that do not collaborate with academic institutions. Sequencing and library preparation costs need to be further reduced, while automation to a greater extent and standardization of sample preparation and bioinformatic analysis will further facilitate

[1] Similarly, advances in diagnostic bacteriology are now obvious due to the metagenomics approach. A number of studies where bacterial pathogens have been successfully detected and identified by metagenomic sequencing are discussed in (Pallen, 2014).

its wider use. Both of these have the potential to offer more coherent results and limited contamination. Microbial contamination - either bacteria or viruses - can take place throughout the process of sample handling, nucleid acids isolation and sequencing. Sources include sample handling in the laboratory or contaminated laboratory reagents and/or nucleic acid extraction kits. There are a number of publications reporting common contaminants and researchers need to keep up to date with this information. It is advised that the same extraction and deep sequencing methods that are used on the clinical samples are also used on negative controls.

The relevance and plausibility of metagenomics findings should always be critically examined, even more so in cases of unlikely findings. The biases of public databases towards cultivable organisms and human pathogens as well as their incompleteness with regards to real biological diversity must be always taken into account. The resulting classification will differ depending on the choice of the database as well as the chosen community profiling method; it has been observed that each method introduces distinct false positives. False positives can be reduced with careful database and method choices, in the end though critical assessment and further validation remains invaluable.

Proving disease causation for metagenomics findings is challenging. The Koch postulates of causation may be modified and adapted in different times, given changes in technology and disease knowledge. Despite the relevance of the newer adapted versions to the molecular era, these cannot be satisfied for viruses in all instances. Viral infection patterns vary along with factors such as genetic susceptibility, age or previous exposure to other agents. Things are easier with acute infectious diseases where the responsible microorganism replicates in the tissue of interest and can be readily identified with traditional methods, there is evidence of an adaptive immune response as well as evident morphological changes consistent with infection. However when classical hallmarks of infection are not present, the pathogenesis mechanism is not direct, the microorganism has latent effects or requires cofactors such as coinfection, confirming causation is more challenging. In such cases, the strength of the epidemiological association in the patients needs to be statistically assessed.

The technology developed by Oxford Nanopore Technologies (ONT) is particularly exciting for medical applications. Their mobile USB-powered single molecule sequencer (MinION) offers the opportunity for patient samples to be directly sequenced in hospitals, reducing the time from sample isolation to diagnosis and subsequent treatment. Furthermore this protability offers new opportunities for developing countries. Using the MinION samples can be sequenced on-site, avoiding long transportation and shipping times. Short response times are often critical in the case of serious outbreaks. This was recently demonstrated with the Ebola virus outbreak in Guinea (Quick *et al.*, 2015) with data released almost instantly, permitting the tracking of the transmission routes in real time. The commercial launch of MinION has not been announced yet and error rates for the nanopore reads are currently higher compared to other sequencing technologies. Still, the promise of low-cost portable sequencer producing very long sequences with low error rates at the point-of-care seem less distant as time goes by. This development has the potential to completely transform infectious disease diagnosis.

Finally, ethical issues arising from metagenomics applications in the clinical setting have not been widely discussed yet. An example of such an issue would be how to proceed with incidental detection of pathogens. Established guidelines for consent and reporting of incidental findings are necessary when handling and analyzing human metagenomic data.

7.3 Final thoughts

This thesis has provided an initial insight in what metagenomic sequencing and methodological work can deliver in a diagnostic capacity. The potential of deep sequencing and bioinformatics solutions have contributed to a sense of great expectations. However in order to reach meaningful conclusions based on the generated data, the choice and suitability of analytic method is essential. It is also of paramount importance that the biases and errors inherent in the sequencing process, the bioinformatics algorithms, the community profiling methods and the public databases are carefully considered.

For this thesis I have contributed novel methodological ideas in the field of method development for community profiling. The main objective was to develop an open-source, sensitive and specific similarity based community profiling method at a high taxonomic resolution, employing the two ideas of parallel MCMC and mixture models. The methodological work has been published in Bioinformatics (Morfopoulou and Plagnol, 2015). metaMix is implemented as an R package, freely available from CRAN, thereby allowing its wider use.

The method owes its conception and development to the nPOD project. Chapter 6 described our involvement in this colloborative effort with the goal to give a definitive answer to whether enteroviruses are implicated in the onset of T1D. The work has not resulted in a positive result, with the caveats mentioned in the previous chapter complicating the interpretation of the negative result. This work has been written as a negative results paper (Morfopoulou *et al.*, Transcriptome sequencing of nPOD type 1 diabetes pancreatic samples for viral sequence identification: insights from the nPOD-V group), pending the results from the few remaining experiments.

Finally, metaMix has been used as the first step for diagnosing clinical cases of patients with undiagnosed viral encephalitis. Two of these results are included in the thesis and are either published (Brown *et al.*, 2015) or are currently under review (Morfopoulou *et al.*, Novel human coronavirus OC43 in an infant with severe combined immunodeficiency and fatal encephalitis, *The New England Journal of Medicine*).

Appendix A

Observations on the assembly step

A.1 Abundant organisms overwhelm real but weak signal during assembly

The primary goal was to assess to what extent the assembled contigs reflect the real abundance or even presence of a species in a metagenomic sample. Simulations designed to quantify how much of the signal is contained in the assembled contigs were carried out. Default settings were used for Velvet.

The first dataset was generated using MetaSim (Richter *et al.*, 2008) designed to simulate a simple scenario. This comprised two viruses arbitrarily chosen, one regarded to be the "signal" (enterovirus) and one the "noise" (measles virus). The idea was to keep the "signal" genome at a steady coverage of either 1x or 5x, while increasing the coverage of the noise. 50 such mixed datasets were simulated. The proportion of the "signal" contained in the assembled contigs reduces as the coverage depth for the noise increases, as demonstrated in Figures A.1 and A.2.

However a typical scenario would entail greater complexity such as detecting rare sequences in an ocean of possibly irrelevant reads. A dataset generated from a real clinical serum sample of high complexity was spiked-in bioinformatically with a low viral signal. The coverage for the spiked-in signal fluctuated from 1x to 10x (50 simulations). Different assemblies were generated, using all non host reads or alternatively using all reads that showed some similarity to viral proteins in RefSeq. In the spiked-in dataset, as the coverage for the true signal reduces, fewer "signal" reads are incorporated into contigs successfully, plotted in Figure A.3.

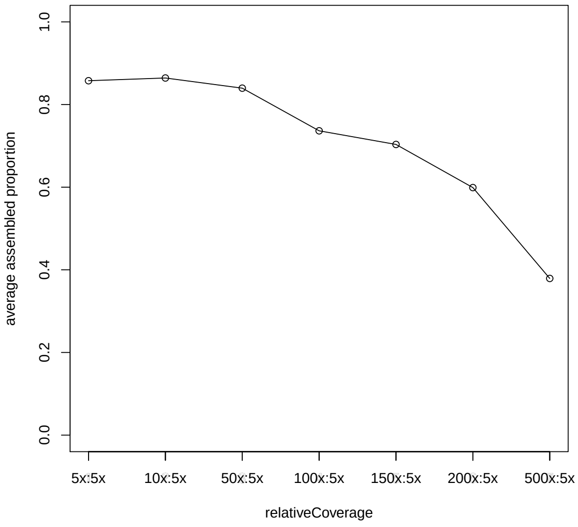


Figure A.1: Simulated 2 genomes dataset: % of "signal" genome contained in assembly, coverage depth at 5x. Coverage depth of noise at 5x-500x. Results based on 50 simulations.

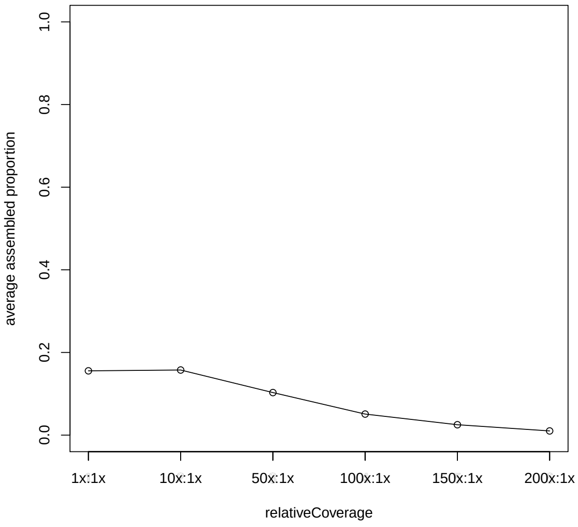


Figure A.2: Simulated 2 genomes dataset: % of "signal" genome contained in assembly, coverage depth at 1x. Coverage depth of noise at 1x-200x. Results based on 50 simulations.

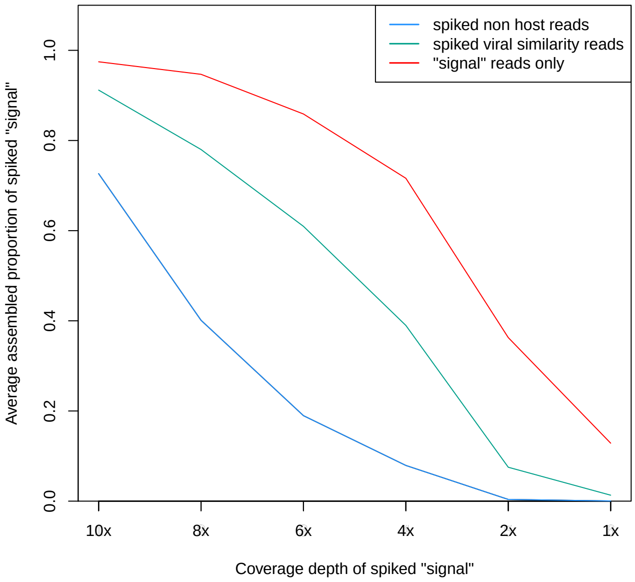


Figure A.3: Spiked-in dataset: % of "signal" genome contained in assembly, coverage depth from 1x to 10x. Results based on 50 simulations.

An alternative approach was borrowed by *de novo* transcriptome assembly (Schulz *et al.*, 2012), (Surget-Groba and Montoya-Burgos, 2010). The main idea is to merge multiple assemblies resulting from different *k*-mer values into one. The motivation behind this approach is to detect differently expressed genes by merging sensitive assemblies with specific ones. It has been shown that longer *k*-values perform best on high expression genes, but poorly on low expression genes. An obvious disadvantage of this approach is that short *k*-mer assemblies may introduce misassemblies.

For metagenomics assembly, one may use an array of *k*-values to partition the metagenomic sequence mix into species bins and then merge the assemblies together. Applying this to the spiked-in dataset with the signal at 1x coverage, 6% of the genome was recovered. The obvious conclusion is that when the coverage is low, the assembly step is going to yield unsatisfactory results despite parameter fine-tuning.

A.2 Unassembled reads are necessary for rare signal detection

An accidental discovery strengthened the argument for using all available reads, including the unassembled ones for species identification, especially when the goal is to detect potential rare pathogens. During the early stages of the project that will be discussed in chapter 6, RNA-seq data from a human post-mortem pancreatic sample were analysed using the assembly-based pipeline This analysis resulted in the identification of two short contigs (200-300bp) originating from a Murine Leukemia virus (MuLV). Upon annotation of the unassembled reads, 50 additional reads mapped confidently to the MuLV genome (Figure A.4) providing further support to the presence of the species in the sample.

It is not difficult to imagine a scenario where due to low coverage it is not possible to obtain any contigs, even after careful and iterative assembly. Low coverage depth for pathogens of interest could for example be encountered when dealing with degraded tissue samples, or when sample collection has happened long after the infection has taken place. Relying only on assembly results would lead in such an instance to miss the opportunity to observe the signal in the sample.

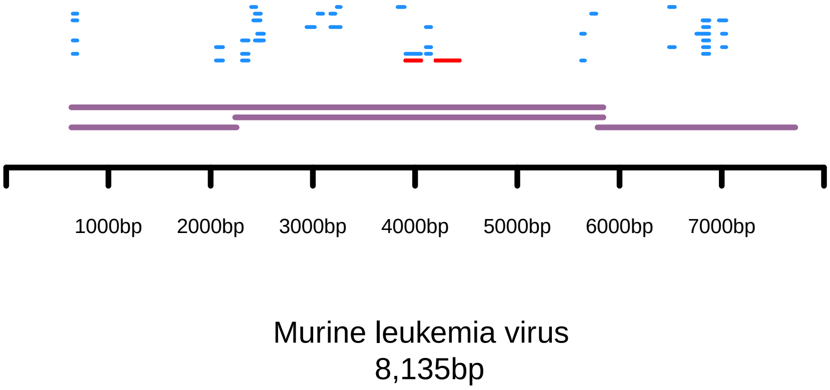


Figure A.4: Alignment view of contigs (red) and reads (blue) against the MuLV genome.

Appendix B

Supplementary Tables

Table B1: simHC FAMeS dataset - predicted species and number of reads assigned to these by metaMix and Pathoscope.

			metaMix	**Pathoscope**	
Taxon	**Scientific**	**True Read**	**Assigned**	**Posterior**	**Final Best Hit**
identifier	**Name**	**Counts**	**Reads**	**Probability**	**Read Numbers**
339671	*Actinobacillus succinogenes 130Z*	483	474	1	135.32
187272	*Alkalilimnicola ehrlichei MLHE-1*	829	798	1	278.52
293826	*Alkaliphillus metalliredigenes UNDEF*	1091	982	1	590.12
240292	*Anabaena variabilis ATCC 29413*	1703	1686	1	507.02
290397	*Anaeromyxobacter dehalogenans 2CP-C*	1273	1263	1	368.52
290399	*Arthrobacter sp. FB24*	1211	1181	1	358.10
322710	***Azotobacter vinelandii AvOP***	1311	1257 1	1	420.53 [1]
315749	*Bacillus cereus NVH391-98*	1000	844	1	242.15
205913	*Bifidobacterium longum DJO10A*	610	557	1	156.58
288000	*Bradyrhizobium sp. BTAi1*	2127	2060	1	731.57
321955	*Brevibacterium linens BL2*	1088	1185	1	565.65
339670	*Burkholderia ambifaria AMMD*	1937	1877	1	706.35
331271	*Burkholderia cenocepacia AU 1054*	1791	2174	1	1256.00
331272	*Burkholderia cenocepacia HI2424*	2045	1656	1	1206.98
269483	***Burkholderia sp. sp.strain 383***	2191	2215 2	1	1009.3 [2]
269482	*Burkholderia vietnamiensis G4*	2083	1989	1	747.98
266265	*Burkholderia xenovorans LB400*	2384	2335	1	799.70
351627	*Caldicellulosiruptor accharolyticus UNDEF*	658	640	1	196.20
290315	*Chlorobium limicola DSMZ 245(T)*	671	638	1	594.15
290317	*Chlorobium phaeobacteroides DSM 266*	719	705	1	631.02

The organisms in bold font are the ones for which metaMix or Pathoscope (or both) identified as present a different strain of the same species or a different species of the same genus.

Table B1 continued: simHC FAMeS dataset - predicted species and number of reads assigned to these by metaMix and Pathoscope .

		metaMix		Pathoscope	
Taxon	**Scientific**	**True Read**	**Assigned**	**Posterior**	**Final Best Hit**
identifier	**Name**	**Counts**	**Reads**	**Probability**	**Read Numbers**
290318	*Chlorobium vvibrioforme f. thiosulfatophilum DSMZ 265(T)*	534	543	1	450.97
324602	***Chloroflexus aurantiacus J-10-fl***	1277	1210 [3]	1	NA
290398	*Chromohalobacter salexigens DSM3043*	888	859	1	284.42
290402	*Clostridium beijerincki NCIMB 8052*	1411	1317	1	453.28
203119	*Clostridium thermocellum ATCC 27405*	932	875	1	338.98
165597	*Crocosphaera watsonii WH 8501*	1593	2897	1	1146.68
269798	*Cytophaga hutchinsonii ATCC 33406*	1161	1102	1	755.40
159087	*Dechloromonas aromatica RCB*	1132	1104	1	421.18
319795	*Deinococcus geothermalis DSM11300*	809	833	1	225.65
272564	*Desulfitobacterium hafniense DCB-2*	1486	1333	1	518.28
207559	*Desulfovibrio desulfuricans G20*	919	858	1	410.38
269484	*Ehrlichia canis Jake*	283	206	1	142.05
332415	*Ehrlichia chaffeensis sapulpa*	255	296	1	201.12
333849	*Enterococcus faecium DO*	676	358	1	124.05
262543	*Exiguobacterium UNDEF 255-15*	788	733	1	275.80
333146	*Ferroplasma acidarmanus fer1*	471	414	1	270.32
106370	*Frankia sp. CcI3*	1334	1278	1	601.87
298653	*Frankia sp. EAN1pec*	2248	2176	1	932.93
269799	*Geobacter metallireducens GS-15*	1025	1035	1	476.35
205914	*Haemophilus somnus 129PT*	513	497	1	148.72
290400	*Jannaschia sp. CCS1*	1148	1037	1	448.28
266940	*Kineococcus radiotolerans SRS30216*	1187	1190	1	440.92
387344	*Lactobacillus brevis ATCC 367*	445	395	1	172.60
321967	*Lactobacillus casei ATCC 334*	648	606	1	235.95
321956	***Lactobacillus delbrueckii bulgaricus ATCC BAA-365***	391	322	1	84.4 [4]
324831	*Lactobacillus gasseri ATCC 33323*	551	555	1	182.17
272622	*Lactococcus lactis cremoris SK11*	584	524	1	155.73
203120	*Leuconostoc mesenteroides mesenteroides ATCC 8293*	473	471	1	155.12
156889	*Magnetococcus sp. MC-1*	1153	999	1	761.50
351348	*Marinobacter aquaeolei VT8*	1164	1127	1	449.60
266779	*Mesorhizobium sp. BNC1*	1289	1244	1	478.72
259564	*Methanococcoides burtonii DSM6242*	663	614	1	329.70

The organisms in bold font are the ones for which metaMix or Pathoscope (or both) identified as present a different strain of the same species or a different species of the same genus.

Table B1 continued: simHC FAMeS dataset - predicted species and number of reads assigned to these by metaMix and Pathoscope .

		metaMix		Pathoscope	
Taxon identifier	Scientific Name	True Read Counts	Assigned Reads	Posterior Probability	Final Best Hit Read Numbers
269797	*Methanosarcina barkeri Fusaro*	1213	1192	1	509.47
323259	*Methanospirillum hungatei JF-1*	919	817	1	589.37
265072	*Methylobacillus flagellatus strain KT*	687	606	1	252.75
264732	*Moorella thermoacetica ATCC 39073*	1426	696	1	352.22
323097	*Nitrobacter hamburgensis UNDEF*	1272	1308	1	672.03
323098	*Nitrobacter winogradskyi Nb-255*	857	727	1	446.38
323261	*Nitrosococcus oceani UNDEF*	868	809	1	304.95
335283	*Nitrosomonas eutropha C71*	649	596	1	245.85
323848	*Nitrosospira multiformis ATCC 25196*	814	757	1	375.75
196162	*Nocardioides sp. JS614*	1337	1333	1	449.52
279238	*Novosphingobium aromaticivorans DSM 12444 (F199)*	1093	1054	1	440.17
203123	*Oenococcus oeni PSU-1*	422	406	1	216.73
318586	*Paracoccus denitrificans PD1222*	1362	1497	1	498.58
278197	*Pediococcus pentosaceus ATCC 25745*	456	451	1	157.27
338963	*Pelobacter carbinolicus DSM 2380*	896	783	1	480.70
338966	*Pelobacter propionicus DSM 2379*	1145	1093	1	531.03
319225	*Pelodictyon luteolum UNDEF*	581	550	1	493.32
324925	*Pelodictyon phaeoclathratiforme BU-1 (DSMZ 5477(T))*	703	676	1	573.25
296591	*Polaromonas sp. JS666*	1489	1468	1	561.72
74546	*Prochlorococcus marinus str. MIT 9312*	404	392	1	124.35
59920	*Prochlorococcus sp. NATL2A*	480	398	1	146.95
290512	*Prosthecochloris aestuarii SK413/DSMZ 271(t)*	692	633	1	481.98
331678	*Prosthecochloris sp. BS1*	1082	769	1	488.95
342610	*Pseudoalteromonas atlantica T6c*	1301	1233	1	474.07
205922	*Pseudomonas fluorescens PfO-1*	1587	1544	1	502.55
351746	*Pseudomonas putida F1*	1528	1542	1	416.47
205918	*Pseudomonas syringae B728a*	1545	1455	1	462.07
259536	*Psychrobacter arcticum 273-4*	623	555	1	259.20
335284	*Psychrobacter cryopegella UNDEF*	793	785	1	393.28
272943	*Rhodobacter sphaeroides 2.4.1*	1119	1130	1	408.28
338969	*Rhodoferax ferrireducens UNDEF*	1276	1258	1	485.52
316055	*Rhodopseudomonas palustris BisA53*	1392	1333	1	760.55

The organisms in bold font are the ones for which metaMix or Pathoscope (or both) identified as present a different strain of the same species or a different species of the same genus.

Table B1 continued: simHC FAMeS dataset - predicted species and number of reads assigned . to these by metaMix and Pathoscope

			metaMix	Pathoscope	
Taxon identifier	**Scientific Name**	**True Read Counts**	**Assigned Reads**	**Posterior Probability**	**Final Best Hit Read Numbers**
316056	*Rhodopseudomonas palustris BisB18*	1348	1344	1	760.72
316057	*Rhodopseudomonas palustris BisB5*	1200	1200	1	805.37
316058	*Rhodopseudomonas palustris HaA2*	1339	1586	1	940.77
269796	*Rhodospirillum rubrum ATCC 11170*	1062	983	1	320.37
266117	*Rubrobacter xylanophilus DSM 9941*	799	758	1	307.28
203122	*Saccharophagus degradans 2-40*	1324	1174	1	796.72
326297	*Shewanella amazonensis SB2B*	1055	953	1	433.48
325240	*Shewanella baltica OS155*	1313	1384	1	678.57
318167	*Shewanella frigidimarina NCMB400*	1257	1133	1	429.40
319224	*Shewanella putefaciens UNDEF*	1153	1438	1	750.38
94122	*Shewanella sp. ANA-3*	1279	1298	1	688.37
60481	*Shewanella sp. MR-7*	1177	1225	1	645.57
323850	*Shewanella sp. PV-4*	1165	1106	1	487.83
351745	*Shewanella sp. W3-18-1*	1214	1015	1	670.82
292414	*Silicibacter sp. TM1040*	1065	1030	1	373.45
317655	*Sphingopyxis alaskensis RB2256*	846	807	1	275.92
286604	*Streptococcus suis 89/1591*	490	1017	1	253.07
322159	*Streptococcus thermophilus LMD-9*	501	290	1	78.97
1140	*Synechococcus sp. PCC 7942 (elongatus)*	646	634	1	200.42
335543	*Syntrophobacter fumaroxidans MPOB*	1181	1067	1	642.47
335541	*Syntrophomonas wolfei Goettingen*	708	649	1	385.48
340099	***Thermoanaerobacter ethanolicus 39E***	570	588 [5]	1	13.05 [6]
269800	*Thermobifida fusca YX*	930	915	1	343.93
292415	*Thiobacillus denitrificans ATCC 25259*	741	742	1	242.93
317025	*Thiomicrospira crunogena XCL-2*	603	563	1	302.28
326298	*Thiomicrospira denitrificans ATCC 33889*	490	503	1	198.52
203124	*Trichodesmium erythraeum IMS101*	2051	1904	1	1156.00
155920	***Xylella fastidiosa Ann-1***	627	NA [7]	NA	NA [7]
155919	*Xylella fastidiosa Dixon*	1303	1343	1	598.32

The organisms in bold font are the ones for which metaMix or Pathoscope (or both) identified as present a different strain of the same species or a different species of the same genus.

[1] Identified Azotobacter vinelandii CA instead (taxon id:1283330).

[2] Burkholderia lata (taxon id: 482957, which includes Burkholderia sp. 383).

[3] Chloroflexus sp. Y-400-fl 1 (taxon id: 480224).

[4] Lactobacillus delbrueckii subsp. Bulgaricus 2038 (taxon id: 353496).

[5] Thermoanaerobacter brockii subsp. finnii Ako-1 (taxon id: 509193).

[6] Thermoanaerobacterium thermosaccharolyticum DSM 571 (taxon id: 580327).

[7] For 155920, the fasta file supposed to contain the reads from it, contained reads from the closely related strain 155919. Therefore even though the community theoretically consists of 113 species, the dataset has reads from 112 organisms.

Table B2: simHC FAMeS dataset - False Positives for metaMix and Pathoscope.

metaMix

Taxon identifier	Scientific Name	Assigned Reads	Posterior Probability
NA	unknown	2284	1
395019	*Burkholderia multivorans ATCC 17616*	131	1
395960	*Rhodopseudomonas palustris TIE-1*	102	1
742013	*Delftia sp. Cs1-4*	98	0.9658119658
866768	*Escherichia coli 'BL21-Gold(DE3)pLysS AG'*	60	0.9252136752

Pathoscope

Taxon identifier	Scientific Name	Final Best Hit Read Numbers
395019	*Burkholderia multivorans ATCC 17616*	687.33
416344	*Burkholderia sp. KJ006*	501.90
407976	*Shewanella baltica OS223*	458.13
335659	*Bradyrhizobium sp. S23321*	295.43
1196325	*Pseudomonas putida DOT-T1E*	289.97
349102	*Rhodobacter sphaeroides ATCC 17025*	242.95
1345695	*Clostridium saccharobutylicum DSM 13864*	164.12
1155766	*Enterococcus faecium Aus0004*	160.72
999541	*Burkholderia gladioli BSR3*	153.08
720554	*Clostridium clariflavum DSM 19732*	105.78
690566	*Sphingobium chlorophenolicum L-1*	100.63
396588	*Thioalkalivibrio sulfidophilus HL-EbGr7*	92.38
65393	*Cyanothece sp. PCC 7424*	89.22
349124	*Halorhodospira halophila SL1*	86.17
526225	*Geodermatophilus obscurus DSM 43160*	83.98
404589	*Anaeromyxobacter sp. Fw109-5*	79.00
497965	*Cyanothece sp. PCC 7822*	78.63
867904	*Methanomethylovorans hollandica DSM 15978*	73.87
483219	*Myxococcus fulvus HW-1*	55.53
266264	*Cupriavidus metallidurans CH34*	53.73
272568	*Gluconacetobacter diazotrophicus PAl 5*	52.30
868597	*Stenotrophomonas maltophilia JV3*	27.77
572480	*Arcobacter nitrofigilis DSM 7299*	26.52
517418	*Chloroherpeton thalassium ATCC 35110*	26.00
1128398	*Clostridium acidurici 9a*	25.37

Table B2 continued: simHC FAMeS dataset - False Positives for metaMix and Pathoscope.

Pathoscope

Taxon identifier	**Scientific Name**	**Final Best Hit Read Numbers**
436717	*Acinetobacter oleivorans DR1*	25.18
929556	*Solitalea canadensis DSM 3403*	24.92
583345	*Methylotenera mobilis JLW8*	23.65
176299	*Agrobacterium fabrum str. C58*	23.27
222523	*Bacillus cereus ATCC 10987*	21.35
1036172	*Arcobacter butzleri 7h1h*	20.40
880070	*Cyclobacterium marinum DSM 745*	19.37
880071	*Flexibacter litoralis DSM 6794*	19.08
762903	*Pedobacter saltans DSM 12145*	16.93
288681	*Bacillus cereus E33L*	16.02
651182	*Desulfobacula toluolica Tol2*	15.83
557599	*Mycobacterium kansasii ATCC 12478*	15.17
866536	*Belliella baltica DSM 15883*	14.88
760192	*Haliscomenobacter hydrossis DSM 1100*	12.38
755732	*Fluviicola taffensis DSM 16823*	11.58
926556	*Echinicola vietnamensis DSM 17526*	10.87
748449	*Halobacteroides halobius DSM 5150*	10.48
1096996	*Acinetobacter baumannii BJAB0715*	10.23
1094466	*Flavobacterium indicum GPTSA100-9*	10.03
110662	*Synechococcus sp. CC9605*	9.88
1041826	*Flavobacterium columnare ATCC 49512*	9.58
411154	*Gramella forsetii KT0803*	8.33
694427	*Paludibacter propionicigenes WB4*	7.68
746697	*Aequorivita sublithincola DSM 14238*	7.53
458233	*Macrococcus caseolyticus JCSC5402*	3.55
867902	*Ornithobacterium rhinotracheale DSM 15997*	2.95
347256	*Mycoplasma hominis ATCC 23114*	2.80
943945	*Mycoplasma fermentans M64*	2.70
515635	*Dictyoglomus turgidum DSM 6724*	2.45

Table B3: Clinical Case 1 - Pathoscope summary.

Taxon identifier	Scientific Name	Final Best Hit Read Numbers
374840	*Enterobacteria phage phiX174 sensu lato*	65327
9606	*Homo sapiens*	554
133448	*Citrobacter youngae*	169
13690	*Sphingobium yanoikuyae*	135
28090	*Acinetobacter lwoffii*	126
469	*Acinetobacter*	123
56946	*Afipia broomeae*	77
409438	*Escherichia coli SE11*	49
645687	*Astrovirus VA1*	46
199310	*Escherichia coli CFT073*	35
1747	*Propionibacterium acnes*	35
1282	*Staphylococcus epidermidis*	10
28211	*Alphaproteobacteria*	10
28037	*Streptococcus mitis*	8
562	*Escherichia coli*	8
509173	*Acinetobacter baumannii AYE*	7
41297	*Sphingomonadaceae*	6
40214	*Acinetobacter johnsonii*	6
29391	*Gemella morbillorum*	5
76122	*Alloprevotella tannerae*	4
652103	*Rhodopseudomonas palustris DX-1*	2
268747	*Prochlorococcus phage P-SSM4*	2

Table B4: Clinical Case 2 - Pathoscope summary (thetaPrior $\in (10, 100)$.

Taxon identifier	Scientific Name	Final Best Hit Read Numbers
31631	*Human coronavirus OC43*	996661
9606	*Homo sapiens*	25036
627439	*Human enteric coronavirus strain 4408*	12498
47229	*Massilia timonae*	538
10090	*Mus musculus*	477
85698	*Achromobacter xylosoxidans*	282
56946	*Afipia broomeae*	119
11128	*Bovine coronavirus*	113
47671	*Lautropia mirabilis*	89
133448	*Citrobacter youngae*	65
509173	*Acinetobacter baumannii AYE*	62
258	*Sphingobacterium spiritivorum*	58
13690	*Sphingobium yanoikuyae*	46
72556	*Achromobacter piechaudii*	43
488	*Neisseria mucosa*	38
158836	*Enterobacter hormaechei*	37
1747	*Propionibacterium acnes*	32
816	*Bacteroides*	29
29430	*Acinetobacter haemolyticus*	22
43767	*Rhodococcus equi*	19
847	*Oxalobacter formigenes*	18
194702	*Cardiobacterium valvarum*	15
250	*Chryseobacterium gleum*	13
469	*Acinetobacter*	12
471	*Acinetobacter calcoaceticus*	12
618	*Serratia odorifera*	12
1034	*Afipia clevelandensis*	10
486	*Neisseria lactamica*	9
502105	*Bovine respiratory coronavirus*	9
729	*Haemophilus parainfluenzae*	8
28037	*Streptococcus mitis*	8
38284	*Corynebacterium accolens*	8
502	*Kingella denitrificans*	7
2718	*Cardiobacterium hominis*	7
40214	*Acinetobacter johnsonii*	7
199310	*Escherichia coli CFT073*	7

Table B4 continued: Clinical Case 2 - Pathoscope summary (thetaPrior $\in (10, 100)$.

Taxon identifier	Scientific Name	Final Best Hit Read Numbers
489	*Neisseria polysaccharea*	6
1282	*Staphylococcus epidermidis*	6
1827	*Rhodococcus*	6
29466	*Veillonella parvula*	5
511145	*Escherichia coli str. K-12 substr. MG1655*	5
1743	*Propionibacterium*	3
648	*Aeromonas caviae*	2
1270	*Micrococcus luteus*	2
331111	*Escherichia coli E24377A*	2
409438	*Escherichia coli SE11*	2
557600	*Acinetobacter baumannii AB307-0294*	2
1292	*Staphylococcus warneri*	1
502108	*Bovine respiratory coronavirus AH187*	1
696748	*Actinobacillus suis H91-0380*	1
698737	*Staphylococcus lugdunensis HKU09-01*	1

Table B5: Mutations in the coronavirus sequence in the brain compared to the GenBank KP198610 genotype E sequence.

POS	REF	ALT	DEPTH	EFFECT	Nucleotide	Amino Acid	Gene
471	C	T	DP=5927	missense	Cat/Tat	p.His88Tyr/c.262C>T	AJC98124.1
2462	C	A	DP=2672	missense	ttC/ttA	p.Phe751Leu/c.2253C>A	AJC98124.1
3202	C	T	DP=2087	missense	gCt/gTt	p.Ala998Val/c.2993C>T	AJC98124.1
3684	A	G	DP=1064	missense	Aag/Gag	p.Lys1159Glu/c.3475A>G	AJC98124.1
4129	G	A	DP=945	missense	gGt/gAt	p.Gly1307Asp/c.3920G>A	AJC98124.1
5753	G	T	DP=474	missense	aaG/aaT	p.Lys1848Asn/c.5544G>T	AJC98124.1
6589	T	C	DP=273	missense	gTt/gCt	p.Val2127Ala/c.6380T>C	AJC98124.1
8089	T	G	DP=104	missense	tTt/tGt	p.Phe2627Cys/c.7880T>G	AJC98124.1
8463	G	A	DP=157	missense	Gtt/Att	p.Val2752Ile/c.8254G>A	AJC98124.1
10864	A	G	DP=191	missense	cAc/cGc	p.His3552Arg/c.10655A>G	AJC98124.1
11287	G	A	DP=90	missense	cGt/cAt	p.Arg3693His/c.11078G>A	AJC98124.1
11919	A	G	DP=108	missense	Agc/Ggc	p.Ser3904Gly/c.11710A>G	AJC98124.1
12547	A	G	DP=126	missense	tAt/tGt	p.Tyr4113Cys/c.12338A>G	AJC98124.1
12738	C	T	DP=141	missense	Ctt/Ttt	p.Leu4177Phe/c.12529C>T	AJC98124.1
13757	G	A	DP=142	missense	Gta/Ata	p.Val4517Ile/c.13549G>A	AJC98124.1
20009	C	G	DP=143	missense	Cag/Gag	p.Gln6601Glu/c.19801C>G	AJC98124.1
20207	G	T	DP=213	missense	Gat/Tat	p.Asp6667Tyr/c.19999G>T	AJC98124.1
22037	A	G	DP=331	missense	Aaa/Gaa	p.Lys178Glu/c.532A>G	ns2a
22537	T	C	DP=262	missense	Tcc/Ccc	p.Ser62Pro/c.184T>C	HE
22577	G	A	DP=249	missense	gGc/gAc	p.Gly75Asp/c.224G>A	HE
22829	T	C	DP=316	missense	aTa/aCa	p.Ile159Thr/c.476T>C	HE
22850	A	G	DP=337	missense	aAt/aGt	p.Asn166Ser/c.497A>G	HE
22871	C	T	DP=340	missense	gCt/gTt	p.Ala173Val/c.518C>T	HE
22886	G	A	DP=342	missense	cGa/cAa	p.Arg178Gln/c.533G>A	HE
23580	G	T	DP=329	missense	ttG/ttT	p.Leu409Phe/c.1227G>T	HE
23736	A	G	DP=375	missense	Ata/Gta	p.Ile33Val/c.97A>G	S
23996	T	A	DP=342	missense	gaT/gaA	p.Asp119Glu/c.357T>A	S
24170	T	A	DP=481	missense	caT/caA	p.His177Gln/c.531T>A	S
24432	G	A	DP=404	missense	Gat/Aat	p.Asp265Asn/c.793G>A	S
24781	C	T	DP=581	missense	gCa/gTa	p.Ala381Val/c.1142C>T	S
25077	C	T	DP=822	missense	Ctt/Ttt	p.Leu480Phe/c.1438C>T	S
25608	A	T	DP=637	missense	Aac/Tac	p.Asn657Tyr/c.1969A>T	S
25699	A	G	DP=450	missense	cAt/cGt	p.His687Arg/c.2060A>G	S
25714	A	T	DP=514	missense	tAt/tTt	p.Tyr692Phe/c.2075A>T	S
25789	C	T	DP=639	missense	aCa/aTa	p.Thr717Ile/c.2150C>T	S

Table B5: Mutations in the coronavirus sequence in the brain compared to the GenBank KP198610 genotype E sequence.

POS	REF	ALT	DEPTH	EFFECT	Nucleotide	Amino Acid	Gene
25924	A	C	DP=796	missense	aAc/aCc	p.Asn762Thr/c.2285A>C	S
26152	T	C	DP=1030	missense	tTa/tCa	p.Leu838Ser/c.2513T>C	S
26328	C	T	DP=645	missense	Cct/Tct	p.Pro897Ser/c.2689C>T	S
26434	A	C	DP=887	missense	gAg/gCg	p.Glu932Ala/c.2795A>C	S
27153	C	A	DP=1402	missense	Cct/Act	p.Pro1172Thr/c.3514C>A	S
27348	C	T	DP=1338	missense	Ccc/Tcc	p.Pro1237Ser/c.3709C>T	S
27444	T	C	DP=1647	missense	Ttc/Ctc	p.Phe1269Leu/c.3805T>C	S
28225	G	T	DP=4918	missense	Gta/Tta	p.Val32Leu/c.94G>T	E
28664	C	T	DP=7144	missense	Ctc/Ttc	p.Leu89Phe/c.265C>T	M
29265	C	T	DP=7455	missense	tCa/tTa	p.Ser55Leu/c.164C>T	N
29727	C	T	DP=7585	missense	tCt/tTt	p.Ser209Phe/c.626C>T	N
29787	C	T	DP=7445	missense	aCa/aTa	p.Thr229Ile/c.686C>T	N
287	A	G	DP=7745	silent	gaA/gaG	p.Glu26Glu/c.78A>G	AJC98124.1
1334	A	C	DP=5910	silent	ggA/ggC	p.Gly375Gly/c.1125A>C	AJC98124.1
2813	C	T	DP=2659	silent	agC/agT	p.Ser868Ser/c.2604C>T	AJC98124.1
4574	C	T	DP=576	silent	taC/taT	p.Tyr1455Tyr/c.4365C>T	AJC98124.1
5528	A	G	DP=358	silent	gaA/gaG	p.Glu1773Glu/c.5319A>G	AJC98124.1
6704	C	T	DP=345	silent	atC/atT	p.Ile2165Ile/c.6495C>T	AJC98124.1
7664	T	C	DP=215	silent	gtT/gtC	p.Val2485Val/c.7455T>C	AJC98124.1
7823	C	T	DP=178	silent	gcC/gcT	p.Ala2538Ala/c.7614C>T	AJC98124.1
9410	C	T	DP=127	silent	aaC/aaT	p.Asn3067Asn/c.9201C>T	AJC98124.1
9668	C	T	DP=82	silent	gtC/gtT	p.Val3153Val/c.9459C>T	AJC98124.1
9914	G	T	DP=82	silent	ccG/ccT	p.Pro3235Pro/c.9705G>T	AJC98124.1
11927	C	T	DP=118	silent	tgC/tgT	p.Cys3906Cys/c.11718C>T	AJC98124.1
13768	T	C	DP=108	silent	ggT/ggC	p.Gly4520Gly/c.13560T>C	AJC98124.1
13783	T	C	DP=100	silent	taT/taC	p.Tyr4525Tyr/c.13575T>C	AJC98124.1
13924	T	C	DP=76	silent	ggT/ggC	p.Gly4572Gly/c.13716T>C	AJC98124.1
14197	G	T	DP=72	silent	acG/acT	p.Thr4663Thr/c.13989G>T	AJC98124.1
14689	C	T	DP=92	silent	acC/acT	p.Thr4827Thr/c.14481C>T	AJC98124.1
14812	C	T	DP=70	silent	ggC/ggT	p.Gly4868Gly/c.14604C>T	AJC98124.1
15733	C	T	DP=143	silent	caC/caT	p.His5175His/c.15525C>T	AJC98124.1
16651	T	C	DP=128	silent	ggT/ggC	p.Gly5481Gly/c.16443T>C	AJC98124.1
17029	C	T	DP=132	silent	caC/caT	p.His5607His/c.16821C>T	AJC98124.1
17107	G	A	DP=103	silent	aaG/aaA	p.Lys5633Lys/c.16899G>A	AJC98124.1

Table B5: Mutations in the coronavirus sequence in the brain compared to the GenBank KP198610 genotype E sequence.

POS	REF	ALT	DEPTH	EFFECT	Nucleotide	Amino Acid	Gene
21991	C	T	DP=247	silent	ccC/ccT	p.Pro162Pro/c.486C>T	ns2a
23046	C	T	DP=227	silent	atC/atT	p.Ile231Ile/c.693C>T	HE
23061	A	T	DP=267	silent	tcA/tcT	p.Ser236Ser/c.708A>T	HE
23518	C	T	DP=413	silent	Cta/Tta	p.Leu389Leu/c.1165C>T	HE
23544	C	T	DP=403	silent	ctC/ctT	p.Leu397Leu/c.1191C>T	HE
24317	C	T	DP=453	silent	acC/acT	p.Thr226Thr/c.678C>T	S
24698	G	T	DP=666	silent	tcG/tcT	p.Ser353Ser/c.1059G>T	S
25199	C	T	DP=835	silent	ggC/ggT	p.Gly520Gly/c.1560C>T	S
25718	T	C	DP=504	silent	caT/caC	p.His693His/c.2079T>C	S
25721	C	T	DP=534	silent	gcC/gcT	p.Ala694Ala/c.2082C>T	S
25769	C	T	DP=535	silent	taC/taT	p.Tyr710Tyr/c.2130C>T	S
25946	C	T	DP=663	silent	atC/atT	p.Ile769Ile/c.2307C>T	S
26009	G	A	DP=677	silent	ttG/ttA	p.Leu790Leu/c.2370G>A	S
26030	T	C	DP=665	silent	taT/taC	p.Tyr797Tyr/c.2391T>C	S
26096	C	T	DP=948	silent	ccC/ccT	p.Pro819Pro/c.2457C>T	S
26216	A	G	DP=1096	silent	gaA/gaG	p.Glu859Glu/c.2577A>G	S
26225	T	C	DP=1162	silent	gaT/gaC	p.Asp862Asp/c.2586T>C	S
26267	T	C	DP=1123	silent	gtT/gtC	p.Val876Val/c.2628T>C	S
26294	T	C	DP=951	silent	ggT/ggC	p.Gly885Gly/c.2655T>C	S
26453	A	T	DP=728	silent	acA/acT	p.Thr938Thr/c.2814A>T	S
26501	C	T	DP=694	silent	ggC/ggT	p.Gly954Gly/c.2862C>T	S
26549	T	C	DP=875	silent	taT/taC	p.Tyr970Tyr/c.2910T>C	S
26564	C	T	DP=922	silent	acC/acT	p.Thr975Thr/c.2925C>T	S
26573	T	C	DP=930	silent	agT/agC	p.Ser978Ser/c.2934T>C	S
26576	A	G	DP=931	silent	ctA/ctG	p.Leu979Leu/c.2937A>G	S
27107	C	T	DP=1560	silent	atC/atT	p.Ile1156Ile/c.3468C>T	S
28651	T	C	DP=6766	silent	ggT/ggC	p.Gly84Gly/c.252T>C	M
28825	T	C	DP=7006	silent	taT/taC	p.Tyr142Tyr/c.426T>C	M
29887	T	C	DP=7411	silent	gtT/gtC	p.Val262Val/c.786T>C	N
30175	G	A	DP=7086	silent	agG/agA	p.Arg358Arg/c.1074G>A	N
26	T	A	DP=1158				
30454	C	T	DP=7269				
30455	G	T	DP=7269				

Table B6: Stage II nPOD-V: metaMix summary profile for sample 6141. The RIN score was 6.7 and the duration of the disease for this case was 28 years.

taxonID	scientName	finalAssignments	poster.prob
374840	*Enterobacteria phage phiX174 sensu lato*	192806	1
unknown	*unknown*	47606	1
133448	*Citrobacter youngae*	11800	1
409438	*Escherichia coli SE11*	8180	1
199310	*Escherichia coli CFT073*	7711	1
2	*Bacteria*	4503	1
158877	*Yokenella regensburgei*	3070	1
13690	*Sphingobium yanoikuyae*	2060	1
69218	*Enterobacter cancerogenus*	1933	1
1747	*Propionibacterium acnes*	1545	1
155864	*Escherichia coli O157:H7 str. EDL933*	1383	1
47229	*Massilia timonae*	1375	1
511145	*Escherichia coli str. K-12 substr. MG1655*	1036	1
56946	*Afipia broomeae*	806	1
331111	*Escherichia coli E24377A*	769	1
618	*Serratia odorifera*	678	1
72556	*Achromobacter piechaudii*	662	1
85698	*Achromobacter xylosoxidans*	659	1
76832	*Myroides odoratimimus*	524	1
655817	*Escherichia coli ABU 83972*	412	1
158836	*Enterobacter hormaechei*	398	0.99
562	*Escherichia coli*	381	1
258	*Sphingobacterium spiritivorum*	361	1
1282	*Staphylococcus epidermidis*	341	1
1035	*Afipia felis*	320	1
847	*Oxalobacter formigenes*	296	1
76831	*Myroides*	279	1
9606	*Homo sapiens*	266	1
204525	*Roseomonas cervicalis*	245	0.99
1034	*Afipia clevelandensis*	223	1
469	*Acinetobacter*	218	1
28211	*Alphaproteobacteria*	198	1
1270	*Micrococcus luteus*	194	1
47671	*Lautropia mirabilis*	191	0.99
250	*Chryseobacterium gleum*	187	1

Table B6 continued: Stage II nPOD-V: metaMix summary profile for sample 6141. The RIN score was 6.7 and the duration of the disease for this case was 28 years.

taxonID	scientName	finalAssignments	poster.prob
1833	*Rhodococcus erythropolis*	184	1
40215	*Acinetobacter junii*	177	1
488	*Neisseria mucosa*	172	1
43767	*Rhodococcus hoagii*	149	0.99
225324	*Enhydrobacter aerosaccus*	147	1
40214	*Acinetobacter johnsonii*	145	1
219314	*Aeromicrobium marinum*	144	1
1019	*Capnocytophaga sputigena*	140	1
816	*Bacteroides*	126	1
471	*Acinetobacter calcoaceticus*	115	1
546271	*Selenomonas sputigena ATCC 35185*	114	1
587	*Providencia rettgeri*	113	1
39488	*[Eubacterium] hallii*	96	1
300269	*Shigella sonnei Ss046*	95	1
29388	*Staphylococcus capitis*	90	1
386585	*Escherichia coli O157:H7 str. Sakai*	89	1
636	*Edwardsiella tarda*	89	1
28090	*Acinetobacter lwoffii*	86	1
548476	*Corynebacterium aurimucosum ATCC 700975*	86	0.96
607712	*Neisseria shayeganii*	81	0.98
1303	*Streptococcus oralis*	71	0.92
102862	*Proteus penneri*	70	0.94
103621	*Actinomyces urogenitalis*	69	0.99
126385	*Providencia alcalifaciens*	69	0.97
1351	*Enterococcus faecalis*	69	1
46503	*Parabacteroides merdae*	68	1
163665	*Dysgonomonas mossii*	63	0.96
569	*Hafnia alvei*	62	1
246787	*Bacteroides cellulosilyticus*	61	0.99
55884	*Enterobacteria phage SfV*	54	0.99
1292	*Staphylococcus warneri*	53	1
43675	*Rothia mucilaginosa*	51	0.99
29430	*Acinetobacter haemolyticus*	49	0.92
43768	*Corynebacterium matruchotii*	49	1
1290	*Staphylococcus hominis*	48	0.99

Table B6 continued: Stage II nPOD-V: metaMix summary profile for sample 6141. The RIN score was 6.7 and the duration of the disease for this case was 28 years.

taxonID	**scientName**	**finalAssignments**	**poster.prob**
1343	*Streptococcus vestibularis*	47	0.98
158	*Treponema denticola*	43	0.97
362663	*Escherichia coli 536*	41	1
331112	*Escherichia coli HS*	39	1
100886	*Catenibacterium mitsuokai*	37	1
28037	*Streptococcus mitis*	37	0.99
364106	*Escherichia coli UTI89*	36	1
1305	*Streptococcus sanguinis*	35	0.99
1827	*Rhodococcus*	35	0.92
1743	*Propionibacterium*	34	0.99
194699	*Bordetella phage BPP-1*	29	0.98
90370	*Salmonella enterica subsp. enterica serovar Typhi*	27	1
198214	*Shigella flexneri 2a str. 301*	24	0.98

Table B7: Stage III nPOD-V: metaMix summary profile for sample 6198. The RIN score was 3.1.

taxonID	scientName	finalAssignments	poster.prob
unknown	*unknown*	438481	1
374840	*Enterobacteria phage phiX174 sensu lato*	37139	1
2	*Bacteria*	3087	1
562	*Escherichia coli*	1583	1
47229	*Massilia timonae*	1272	1
1343	*Streptococcus vestibularis*	1222	1
1747	*Propionibacterium acnes*	1152	1
1034	*Afipia clevelandensis*	1093	1
56946	*Afipia broomeae*	1093	1
1035	*Afipia felis*	919	1
13690	*Sphingobium yanoikuyae*	552	1
69218	*Enterobacter cancerogenus*	495	1
199310	*Escherichia coli CFT073*	491	1
72556	*Achromobacter piechaudii*	439	1
409438	*Escherichia coli SE11*	421	1
85698	*Achromobacter xylosoxidans*	407	1
10090	*Mus musculus*	310	1
158877	*Yokenella regensburgei*	280	1
133448	*Citrobacter youngae*	261	1
9606	*Homo sapiens*	249	1
1304	*Streptococcus salivarius*	246	1
469	*Acinetobacter*	227	1
1033	*Afipia*	224	0.99
40214	*Acinetobacter johnsonii*	196	1
618	*Serratia odorifera*	183	0.97
471	*Acinetobacter calcoaceticus*	144	1
1305	*Streptococcus sanguinis*	105	1
488	*Neisseria mucosa*	95	1
41294	*Bradyrhizobiaceae*	89	1
511145	*Escherichia coli str. K-12 substr. MG1655*	87	1
204525	*Roseomonas cervicalis*	81	0.97
1303	*Streptococcus oralis*	80	0.98
28037	*Streptococcus mitis*	79	1
1301	*Streptococcus*	76	1
1282	*Staphylococcus epidermidis*	70	1

Table B7 continued: Stage III nPOD-V: metaMix summary profile for sample 6198. The RIN score was 3.1.

taxonID	scientName	finalAssignments	poster.prob
40215	*Acinetobacter junii*	69	0.97
28090	*Acinetobacter lwoffii*	68	1
250	*Chryseobacterium gleum*	67	1
47671	*Lautropia mirabilis*	67	1
225324	*Enhydrobacter aerosaccus*	66	1
155864	*Escherichia coli O157:H7 str. EDL933*	65	0.99
267747	*Propionibacterium acnes KPA171202*	64	1
28211	*Alphaproteobacteria*	60	1
762948	*Rothia dentocariosa ATCC 17931*	56	0.98
1270	*Micrococcus luteus*	51	1
553199	*Propionibacterium acnes SK137*	49	0.98
43768	*Corynebacterium matruchotii*	47	1
331111	*Escherichia coli E24377A*	45	1
587	*Providencia rettgeri*	45	0.92
1833	*Rhodococcus erythropolis*	43	0.93
38304	*Corynebacterium tuberculostearicum*	39	0.97
39778	*Veillonella dispar*	29	0.95
470	*Acinetobacter baumannii*	21	0.91

Appendix C

Supplementary Text

Parameter setting for Phylogenetic Tree estimation

The following options were used for the construction of the phylogram: General Time Reversible model, Bootstrap branch support of 100 replicates, optimized invariable sites and across site rate variation. The tree searching operation used was the best of NNI snd SPR, with starting tree BioNJ and optimizing the tree topology.

Bibliography

Altschul, S.F. et al (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Altschul, S.F. et al (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–402.

Ann Yeh, Arlene Collins, Michael Cohen, P.D. and Faden, H. (2004). Detection of Coronavirus in the Central Nervous System of a Child With Acute Disseminated Encephalomyelitis. *Pediatrics*, **113**(1), e73–e76.

Arbour, N. et al (1999). Acute and persistent infection of human neural cell lines by human coronavirus OC43. *Journal of virology*, **73**(4), 3338–3350.

Arbour, N. et al (2000). Neuroinvasion by human respiratory coronaviruses. *Journal of virology*, **74**(19), 8913–8921.

Assiri, A. et al (2013). Hospital outbreak of Middle East respiratory syndrome coronavirus. *The New England Journal of Medicine*, **369**(5), 407–16.

Atkinson, M.a. (2014). Pancreatic biopsies in type 1 diabetes: Revisiting the myth of Pandora's box. *Diabetologia*, **57**(4), 656–659.

Bach, J.F. (2002). The effect of infections on susceptibility to autoimmune and allergic diseases. *The New England Journal of Medicine*, **347**(12), 911–920.

Barzon, L. et al (2013). Next-generation sequencing technologies in diagnostic virology. *Journal of Clinical Virology*, **58**(2), 346–50.

Bentley, D.R. et al (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–9.

Bentley, S.D. and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annual review of genetics*, **38**(13), 771–92.

Bhaduri, A. et al (2012). Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics*, **28**(8), 1174–1175.

Bibby, K. (2013). Metagenomic identification of viral pathogens. *Trends in biotechnology*, **31**(5), 275–9.

Blomström, A.L. et al (2010). Detection of a Novel Astrovirus in Brain Tissue of Mink Suffering from Shaking Mink Syndrome by Use of Viral Metagenomics. *Journal of Clinical Microbiology*, **48**(12), 4392–4396.

Bluestone, J.a. et al (2010). Genetics, pathogenesis and clinical interventions in type1 diabetes. *Nature*, **464**(7293), 1293–1300.

Brady, A. and Salzberg, S.L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, **6**(9), 673–6.

Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, **13**(6), 278–84.

Brooks, S.P. (1998). Markov Chain Monte Carlo Method and Its Application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **47**(1), 69–100.

Brown, J.R. et al (2015). Astrovirus VA1/HMO-C: An Increasingly Recognized Neurotropic Pathogen in Immunocompromised Patients. *Clinical Infectious Diseases*, pages 1–8.

Burrows, M. and Wheeler, D. (1994). A block-sorting lossless data compression algorithm. *Algorithm, Data Compression*, (124), 18.

Butler, J. et al (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, **18**(5), 810–20.

Campbell-Thompson, M. et al (2012). Network for Pancreatic Organ Donors with Diabetes (nPOD): Developing a tissue biobank for type 1 diabetes. *Diabetes/Metabolism Research and Reviews*, **28**(7), 608–617.

Chen, K. and Pachter, L. (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLoS Comput Biol*, **1**(2), e24.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**(4), 327–335.

Chin, C.S. et al (2011). The Origin of the Haitian Cholera Outbreak Strain. *The New England Journal of Medicine*, (364), 33–42.

Chiu, C.Y. (2013). Viral pathogen discovery. *Current Opinion in Microbiology*, **16**(4), 468–78.

Christen, U. and von Herrath, M.G. (2011). Do viral infections protect from or enhance type 1 diabetes and how can we tell the difference? *Cellular and Molecular Immunology*, **8**(3), 193–198.

Cingolani, P. et al (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, **6:2**(June), 1–13.

Clayton, D.G. (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS genetics*, **5**(7), e1000540.

Coppieters, K.T. et al (2012). Virus infections in type 1 diabetes. *Cold Spring Harbor perspectives in medicine*, **2**(1), a007682.

Cotten, M. et al (2014). Deep Sequencing of Norovirus Genomes Defines Evolutionary Patterns in an Urban Tropical Setting. *Journal of Virology*, **88**(19), 11056–11069.

Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, **12**, 138–163.

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

Day, W. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, **49**(4), 461–467.

Delsuc, F. et al (2005). Phylogenomics and the reconstruction of the tree of life. *Nature reviews. Genetics*, **6**(5), 361–375.

Dempster, A. and Laird, N. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.*, **39**(1), 1–38.

Deng, Y.M. et al (2011). Rapid Detection and Subtyping of Human Influenza A Viruses and Reassortants by Pyrosequencing. *PLoS ONE*, **6**(8), e23400.

Depledge, D.P. et al (2011). Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. *PLoS ONE*, **6**(11).

Deschavanne, P.J. et al (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular biology and evolution*, **16**(10), 1391–9.

Desforges, M. et al (2014). Human coronaviruses: Viral and cellular factors involved in neuroinvasiveness and neuropathogenesis. *Virus Research*, **194**, 145–158.

Didelot, X. et al (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature reviews. Genetics*, **13**(9), 601–12.

Diebolt, J. and Robert, C. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society Series B Methodological*, **56**(2), 363–375.

Dotta, F. et al (2007). Coxsackie B4 virus infection of beta cells and natural killer cell insulitis in recent-onset type 1 diabetic patients. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(12), 5115–5120.

Dröge, J. and McHardy, A.C. (2012). Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics*, **13**(6), 646–55.

Earl, D.J. and Deem, M.W. (2005). Parallel tempering: theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, **7**(23), 3910–6.

Eid, J. et al (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, **323**, 133–138.

Erlich, Y. and Mitra, P. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature methods*, **5**(8), 679–682.

Falkow, S. (2004). Molecular Koch's postulates applied to bacterial pathogenicity–a personal recollection 15 years later. *Nature reviews. Microbiology*, **2**(1), 67–72.

Fancello, L. et al (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology*, **434**(2), 162–174.

Favreau, D.J. et al (2009). A human coronavirus OC43 variant harboring persistence-associated mutations in the S glycoprotein differentially induces the unfolded protein response in human neurons as compared to wild-type virus. *Virology*, **395**(2), 255–267.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**(6), 368–376.

Flaherty, P. et al (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Research*, **40**(1), 1–12.

Flicek, P. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, **6**(11).

Foulis, a.K. et al (1987). Aberrant expression of class II major histocompatibility complex molecules by B cells and hyperexpression of class I major histocompatibility complex molecules by insulin containing islets in type 1 (insulin-dependent) diabetes mellitus. *Diabetologia*, **30**(5), 333–343.

Francis, O.E. et al (2013). Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Research*, **23**(10), 1721–9.

Frank, C. et al (2011). Epidemic Profile of Shiga-ToxinProducing Escherichial coli O104:H4 Outbreak in Germany. *The New England Journal of Medicine*, (365), 1771–80.

Fredricks, D.N. and Relman, D.a. (1996). Sequence-based identification of microbial pathogens: A reconsideration of Koch's postulates. *Clinical Microbiology Reviews*, **9**(1), 18–33.

Funchain, P. and Charis, E. (2012). Hunting for cancer in the microbial jungle. *Genome medicine*, **5**(5), 42.

Gale, E.A. (2002). The rise of childhood type 1 diabetes in the 20th century. *Diabetes*, **51**(12), 3353–3361.

Gelman, A. et al (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Giannoukos, G. et al (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology*, **13**(3), R23.

Gilks, W. (1999). *Markov Chain Monte Carlo In Practice*. Chapman and Hall/CRC.

Glaser, C.a. et al (2006). Beyond viruses: clinical profiles and etiologies associated with encephalitis. *Clinical infectious diseases*, **43**(November), 1565–1577.

Gouy, M. et al (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, **27**(2), 221–4.

Graham, R.L. et al (2013). A decade after SARS: strategies for controlling emerging coronaviruses. *Nature reviews. Microbiology*, **11**(12), 836–48.

Granerod, J. et al (2010). Challenge of the unknown: A systematic review of acute encephalitis in non-outbreak situations. *Neurology*, **75**, 924–932.

Green, J. et al (2004). Coxsackie b virus serology and type 1 diabetes mellitus: a systematic review of published case-control studies. *Diabetic Medicine*, **21**(6), 507–514.

Greninger, A.L. et al (2010). A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PloS one*, **5**(10), e13381.

Guindon, S. et al (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**(3), 307–21.

Gull, S. (1988). Bayesian inductive inference and maximum entropy. **31-32**, 53–74.

Handley, S.a. et al (2012). Pathogenic simian immunodeficiency virus infection is associated with expansion of the enteric virome. *Cell*, **151**(2), 253–266.

Harjutsalo, V. (2008). Time trends in the incidence of type 1 diabetes in Finnish children: a cohort study. *The Lancet*, **10**(Icd), 1777–1782.

Hastings, W.K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**(1), 97–109.

Hershey, A.D. and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, pages 645–656.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, **37**(2), 185–194.

Hoeting, J.a. (2002). Methodology for Bayesian Model Averaging : An Update. *International Biometric Conference*, pages 231–240.

Hoeting, J.A. et al (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Holder, M. and Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews. Genetics*, **4**(4), 275–284.

Holley, R.W. et al (1965). Structure of a Ribonucleic Acid. *Science*, **147**(3664), 1462–1465.

Hué, S. et al (2010). Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology*, **7**(1), 111.

Huelsenbeck, J.P. et al (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science (New York, N.Y.)*, **294**(5550), 2310–2314.

Huson, D.H. et al (2007). MEGAN analysis of metagenomic data. *Genome Research*, **17**(3), 377–386.

Jacomy, H. et al (2006). Human coronavirus OC43 infection induces chronic encephalitis leading to disabilities in BALB/C mice. *Virology*, **349**, 335–346.

Jasra, A. et al (2007). On population-based simulation for static inference. *Statistics and Computing*, **17**(3), 263–279.

Jeffreys, H. (1961). *Theory of Probability (3rd Edition)*. Oxford University Press, New York.

Kass, R.R.E. and Raftery, A.E.A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773– 795.

Kislyuk, A. et al (2009). Unsupervised statistical clustering of environmental shotgun sequences. *BMC bioinformatics*, **10**, 316.

Knox, K. et al (2011). No evidence of murine-like gammaretroviruses in CFS patients previously identified as XMRV-infected. *Science*, **333**(6038), 94–97.

Koboldt, D.C. et al (2012). VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. pages 568–576.

Kozarewa, I. et al (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C) -biased genomes. *Nature methods*, **6**(4), 291–295.

Krogvold, L. et al (2014). Pancreatic biopsy by minimal tail resection in live adult patients at the onset of type 1 diabetes: Experiences from the DiViD study. *Diabetologia*, **57**(4), 841–843.

Krogvold, L. et al (2015). Detection of a Low-Grade Enteroviral Infection in the Islets of Langerhans of Living Patients Newly Diagnosed With Type 1 Diabetes. *Diabetes*, **64**(5), 1682–1687.

Kunin, V. et al (2008). A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, **72**(4), 557–78, Table of Contents.

Kuroda, M. et al (2010). Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. *PloS one*, **5**(4), e10256.

Lander, E.S. et al (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

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

Lau, S.K.P. et al (2011). Molecular Epidemiology of Human Coronavirus OC43 Reveals Evolution of Different Genotypes over Time and Recent Emergence of a Novel Genotype due to Natural Recombination. *Journal of Virology*, **85**(21), 11325–11337.

Lecuit, M. and Eloit, M. (2013). The human virome: new tools and concepts. *Trends in Microbiology*, **21**(10), 510–515.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–2993.

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

Li, H. et al (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, R. et al (2009b). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, **25**(15), 1966–1967.

Li, R. et al (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, **20**(2), 265–72.

Lipkin, W.I. (2008). Pathogen Discovery. *PLoS Pathogens*, **4**(4), e1000002.

Lipkin, W.I. (2009). Microbe hunting in the 21st century. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(1), 6–7.

Lipkin, W.I. and Hornig, M. (2015). Diagnostics and Discovery in Viral Central Nervous System Infections. *Brain Pathology*, **25**(5), 600–604.

Liu, J.S. (2001). *Monte Carlo strategies in scientific computing*. Springer.

Liu, L. et al (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, **2012**, 251364.

Lole, K.S. et al (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of virology*, **73**(1), 152–160.

Loman, N.J. et al (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**(5), 434439.

Mackay, D.J.C. (1992). *Bayesian Methods for Adaptive Models*. Ph.D. thesis, California Institute of Technology, Pasadena, CA, USA.

Marchesi, J.R. and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, **3**(1), 31.

Margulies, M. et al (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–80.

Marin, J.M. and Robert, C. (2008). Approximating the marginal likelihood in mixture models. *ArXiv e-prints*.

Marin, J.M. et al (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, **25**.

Markowitz, V.M. et al (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, **40**(D1), D115–D122.

Marra, M.a. et al (2003). The Genome sequence of the SARS-associated coronavirus. *Science (New York, N.Y.)*, **300**(May), 1399–1404.

Masuda, N. et al (1999). Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Research*, **27**(22), 4436–4443.

Matranga, C.B. et al (2014). Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biology*, **15**(11), 519.

Matsen, F.a. (2015). Phylogenetics and the Human Microbiome. *Systematic Biology*, **64**(1), e26–e41.

Mavromatis, K. et al (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods*, **4**(6), 495–500.

McElroy, K. et al (2013). Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias. *BMC genomics*, **14**(1), 501.

McElroy, K. et al (2014). Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microbial informatics and experimentation*, **4**(1), 1.

McHardy, A.C. and Rigoutsos, I. (2007). What's in the mix: phylogenetic classification of metagenome sequence samples. *Current opinion in microbiology*, **10**(5), 499–503.

McHardy, A.C. et al (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, **4**(1), 63–72.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics. Wiley.

McMullan, L.K. et al (2012). A new phlebovirus associated with severe febrile illness in Missouri. *The New England Journal of Medicine*, **367**(9), 834–41.

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

Minot, S. et al (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research*, **21**(10), 1616–25.

Mizuno, C.M. et al (2013). Expanding the marine virosphere using metagenomics. *PLoS Genetics*, **9**(12), e1003987.

Morange, M. (2009). History of Molecular Biology. *Life Sciences*, pages 1–9.

Morfopoulou, S. and Plagnol, V. (2015). Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics*.

Mullis, K.B. and Faloona, F.a. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, **155**, 335–350.

Myers, E.W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, **2**, 275–290.

Naccache, S.N. et al (2013). The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. *Journal of Virology*, **87**(22), 11966–11977.

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), 443–453.

Negredo, A. et al (2011). Discovery of an ebolavirus-like filovirus in europe. *PLoS Pathogens*, **7**(10), e1002304.

Nejentsev, S. et al (2009). Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**(5925), 387–389.

Newton, M.A. and Raftery, A.E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

Ng, T.F.F. et al (2011). Broad Surveys of DNA Viral Diversity Obtained through Viral Metagenomics of Mosquitoes. *PLoS ONE*, **6**(6), e20579.

Oakes, B. et al (2010). Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. *Retrovirology*, pages 1–10.

Oikarinen, S. et al (2011). Enterovirus RNA in blood is linked to the development of type 1 diabetes. *Diabetes*, **60**(1), 276–279.

Oikarinen, S. et al (2014). Virus antibody survey in different european populations indicates risk association between coxsackievirus B1 and type 1 diabetes. *Diabetes*, **63**(2), 655–662.

Ounit, R. et al (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, **16**(1), 236.

Ozsolak, F. and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, **12**(2), 87–98.

Pallen, M.J. (2014). Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*, **141**(14), 1856–1862.

Peiris, J. et al (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet*, **361**, 1319–1325.

Perlman, S. and Netland, J. (2009). Coronaviruses post-SARS: update on replication and pathogenesis. *Nature reviews. Microbiology*, **7**(juNE), 439–450.

Petrosino, J.F. et al (2009). Metagenomic pyrosequencing and microbial identification. *Clinical chemistry*, **55**(5), 856–66.

Pevzner, P.A. et al (2001). An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, **98**(17), 9748–9753.

Pierre J. Talbot, Marc Desforges, E.B. and Jacomy, H. (2011). *Non-Flavivirus Encephalitis*. InTech.

Pihoker, C. et al (2005). Autoantibodies in diabetes. *Diabetes*, **54**(suppl 2), S52–S61.

Polychronakos, C. and Li, Q. (2011). Understanding type 1 diabetes through genetics: advances and prospects. *Nature reviews. Genetics*, **12**(11), 781–92.

Pray, L. (2008). Discovery of DNA Structure and Function : Watson and Crick. *Nature Education*, **1**(1), 1–6.

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

Quail, M.a. et al (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, **13**(1), 341.

Quan, P. et al (2010). Astrovirus encephalitis in boy with x-linked agammaglobulinemia. *Emerging Infectious Diseases*, **16**(6), 918925.

Quiñones Mateu, M.E. et al (2014). Deep sequencing: Becoming a critical tool in clinical virology. *Journal of Clinical Virology*, **61**(1), 9–19.

Quick, J. et al (2015). Real-time sequencing and data release for Ebolavirus genomic surveillance in Guinea. virological.org.

Redondo, M.J. et al (1999). Genetic determination of islet cell autoimmunity in monozygotic twin, dizygotic twin, and non-twin siblings of patients with type 1 diabetes: prospective twin study. *BMJ*, **318**(7185), 698–702.

Richardson, S. et al (2009). The prevalence of enteroviral capsid protein vp1 immunostaining in pancreatic islets in human type 1 diabetes. *Diabetologia*, **52**, 1143–1151. 10.1007/s00125-009-1276-0.

Richardson, S.J. et al (2011). Immunopathology of the human pancreas in type-I diabetes. *Seminars in Immunopathology*, **33**(1), 9–21.

Richardson, S.J. et al (2013). Expression of the enteroviral capsid protein VP1 in the islet cells of patients with type 1 diabetes is associated with induction of protein kinase R and downregulation of Mcl-1. *Diabetologia*, **56**(1), 185–193.

Richardson, S.J. et al (2014a). Detection of enterovirus in the islet cells of patients with type 1 diabetes: What do we learn from immunohistochemistry? Reply to Hansson SF, Korsgren S, Pontén F et al [letter]. *Diabetologia*, **57**(3), 647–649.

Richardson, S.J. et al (2014b). Pancreatic pathology in type 1 diabetes mellitus. *Endocrine Pathology*, **25**(1), 80–92.

Richter, D.C. et al (2008). MetaSim: a sequencing simulator for genomics and metagenomics. *PloS one*, **3**(10), e3373.

Rivers, T.M. (1937). Viruses and Koch's Postulates. *Journal of bacteriology*, **33**(1), 1–12.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Roberts, L. et al (2009). Identification of methods for use of formalin-fixed, paraffin-embedded tissue samples in RNA expression profiling. *Genomics*, **94**(5), 341–348.

Robinson, M.J. et al (2010). Mouse DNA contamination in human tissue tested for XMRV. *Retrovirology*, **7**(1), 108.

Rohde, H. et al (2011). Open-Source Genomic Analysis of Shiga-ToxinProducing. *The New England Journal of Medicine*, (365), 718–24.

Rosseel, T. et al (2014). False-Positive Results in Metagenomic Virus Discovery: A Strong Case for Follow-Up Diagnosis. *Transboundary and Emerging Diseases*, **61**(4), 293–299.

Rothberg, J.M. et al (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**(7356), 348–352.

Saiki, R.K. et al (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)*, **239**(4839), 487–491.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4), 406–425.

Salter, S. et al (2014). Reagent contamination can critically impact sequence-based microbiome analyses. *bioRxiv*, page 10.1101/007187.

Sanger, F. and Coulson, a.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, **94**(3), 441–448.

Sanger, F. et al (1977a). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–7.

Sanger, F. et al (1977b). Nucleotide sequence of bacteriophage [phi]x174 dna. *Nature*, **265**(5596), 687–695.

Santos, F. et al (2011). Metatranscriptomic analysis of extremely halophilic viral communities. *The ISME Journal*, **5**(10), 1621–1633.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**(6), 863–4.

Schroeder, A. et al (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, **7**, 3.

Schulz, M.H. et al (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics.*, **28**(8), 1086–92.

Segata, N. et al (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**(8), 811–814.

Shendure, J. et al (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**(5741), 1728–32.

Sievers, F. et al (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**(1), 539.

Sigurgeirsson, B. et al (2014). Sequencing degraded RNA addressed by 3' tag counting. *PloS one*, **9**(3), e91851.

Silkie, S.S. et al (2008). Reagent decontamination to eliminate false-positives in Escherichia coli qPCR. *Journal of Microbiological Methods*, **72**(3), 275–282.

Simmons, G. et al (2011). Failure to confirm XMRV/MLVs in the blood of patients with chronic fatigue syndrome: a multi-laboratory study. *Science (New York, N.Y.)*, **334**(6057), 814–7.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

St-Jean, J.R. et al (2004). Human respiratory coronavirus OC43: genetic stability and neuroinvasion. *Journal of Virology*, **78**(16), 8824–8834.

Sultan, M. et al (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC genomics*, **15**(1), 675.

Surget-Groba, Y. and Montoya-Burgos, J.I. (2010). Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome research*, **20**(10), 1432–40.

Teeling, H. et al (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics*, **5**, 163.

Thompson, C. et al (2012). Encephalitis in children. *Archives of Disease in Childhood*, **97**(2), 150–161.

Thorsby, E. (1997). Invited anniversary review: HLA associated diseases. *Human Immunology*, **8859**(97).

Tunkel, A.R. et al (2008). The management of encephalitis: clinical practice guidelines by the Infectious Diseases Society of America. *Clinical infectious diseases.*, **47**(3), 303–27.

Tuomilehto, J. et al (1999). Record-high incidence of type i (insulin-dependent) diabetes mellitus in finnish children. *Diabetologia*, **42**, 655–660. 10.1007/s001250051212.

Vijgen, L. et al (2005a). Circulation of genetically distinct contemporary human coronavirus OC43 strains. *Virology*, **337**, 85–92.

Vijgen, L. et al (2005b). Complete Genomic Sequence of Human Coronavirus OC43 : Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event. *Journal of Virology*, **79**(3), 1595–1604.

Virgin, H.W. et al (2009). Redefining chronic viral infection. *Cell*, **138**(1), 30–50.

von Herrath, M.G. et al (2003). Microorganisms and autoimmunity: making the barren field fertile? *Nature reviews. Microbiology*, **1**(2), 151–7.

Waner, J.L. (1994). Mixed viral infections: detection and management. *Clinical microbiology reviews*, **7**(2), 143–51.

Wang, Z. et al (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**(1), 57–63.

Watson, J. and Crick, F.H.F. (1953a). Molecular structure of nucleic acids. *Nature*, **171**(4356), 737–8.

Watson, J.D. and Crick, F.H. (1953b). Genetical implications of the structure of deoxyribonucleic acid.

Willner, D. et al (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*, **4**(10), e7370.

Wilm, A. et al (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, **40**(22), 11189–11201.

Wilson, M.R. et al (2014). Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *The New England Journal of Medicine*, **370**(April 2013), 2408–17.

Wood, D.E. and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, **15**(3), R46.

Xia, L.C. et al (2011). Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, **6**(12), e27992.

Xie, W. et al (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, **60**(2), 150–160.

Xu, J. et al (2005). Detection of severe acute respiratory syndrome coronavirus in the brain: potential role of the chemokine mig in pathogenesis. *Clinical infectious diseases*, **41**, 1089–1096.

Yang, J. et al (2011). Unbiased Parallel Detection of Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach. *Journal of Clinical Microbiology*, **49**(10), 3463–3469.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, **10**(6), 1396–1401.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, **13**(5), 303–314.

Yeung, W.C.G. et al (2011). Enterovirus infection and type 1 diabetes mellitus: systematic review and meta-analysis of observational molecular studies. *Bmj*, **342**(feb03 1), d35—-d35.

Yoon, J.W. et al (1979). Virus-induced diabetes mellitus. *The New England Journal of Medicine*, **300**(21), 1173–1179. PMID: 219345.

Yozwiak, N.L. et al (2012). Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing. *PLoS Neglected Tropical Diseases*, **6**(2), e1485.

Zaki, A.M. et al (2012). Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *The New England Journal of Medicine*, page 121017140031005.

Zerbino, D.R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**(5), 821–829.

Zhang, X.M. et al (1994). Biological and genetic characterization of a hemagglutinating coronavirus isolated from a diarrhoeic child. *J Med Virol*, **44**(2), 152–161.

Zhang, Y. et al (2014). Genotype shift in human coronavirus OC43 and emergence of a novel genotype by natural recombination. *Journal of Infection*, (9).

Zhao, W. et al (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC genomics*, **15**(1), 419.

Ziegler, A.G. and Nepom, G.T. (2010). Prediction and pathogenesis in type 1 diabetes. *Immunity*, **32**(4), 468 – 478.