MONITORING THE HUMAN INFECTOME

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Lance Martin
January 2015

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Steve Quake) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Howard Chang)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Peter Sarnow)

Approved for the University Committee on Graduate Studies.

Abstract

Hello world!

Acknowledgments

This is the acknowledgement!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 History of infectious disease

Infectious diseases have a profound impact on humankind, influencing the course of wars and the fate of nations. Only two centuries ago, infectious diseases were a defining challenge of the human condition. For perspective, consider that George Washington was born in 1732, a time when there was no well-defined concept of infection or immunity, no vaccines, and not effective treatments for infectious diseases. Washinton suffered from smallpox and malaria, wound infections and abscesses, and nursed his brother on a tropical island as he died of tuberculosis [7]. Almost all the major advances in the understanding and control of infectious diseases occurred in the two centuries since the founding of the United States.

Advances began with the first animal-transmission studies conducted soon after the War of 1812. These were followed by the development and improvement of microscopes, which for the first time linked micro-organisms to skin and mucosal diseases. Robert Koch developed unifying principles for infectious disease in the late 1800s, providing criteria to establish a causal link between micro-organism and disease. In the early 20th century, Paul Ehrlich developed anti-infective serums to kill pathogens, which paved the way for the vaccines, antibiotics, and antiviral agents that saved hundreds of millions of lives and extended the human life span.

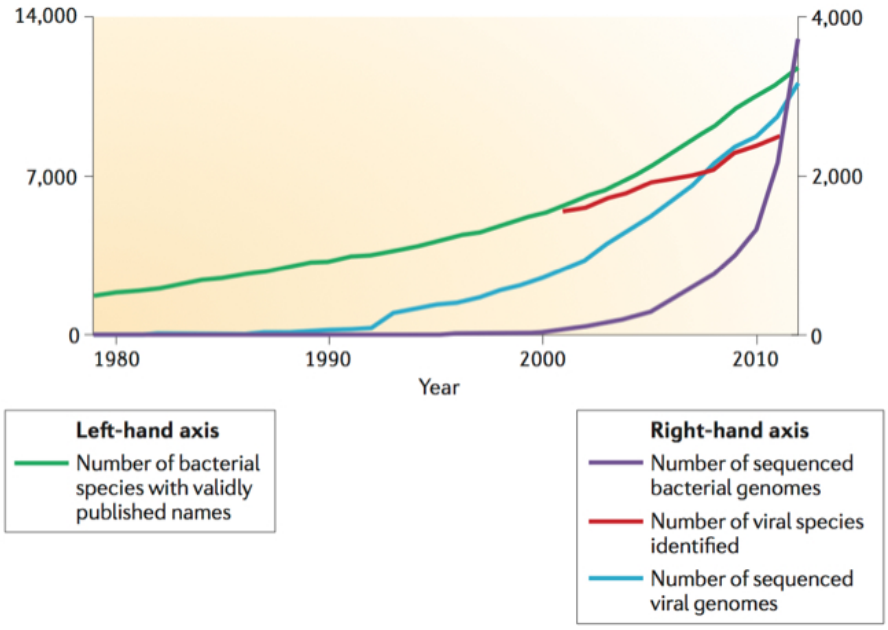

Figure 1.1: Rapid growth in the identification of micro-organisms.

1.2 Technology

1.2.1 Historical perspective

Since the seminal contributions of scientists like Koch and Ehrlich, technology has improved our understanding of infectious disease. Though a cornerstone of microbiology since the nineteenth century, culture fails to cultivate or distinguish many microbes. By 1980, only 1800 validated bacterial species had been published [8]. DNA-based analyses changed the paradigm, as they enabled identification and taxonomic classification of micro-organisms based on genetic material (DNA or RNA).

Hantavirus pulmonary syndrome, an ancient disease caused by a phlebovirus, was discovered unexpectedly in 1993 by the application of a powerful DNA-based assay, polymerase chain reaction (PCR). Less than a year later, PCR-related subtraction techniques solved a century-old mystery of the cause of Kaposiś sarcoma. Since that time, DNA-based analyses have become cheaper and more effective. They have ushered in an era of rapid micro-organism discovery (Figure 1.1) [8].

1.2.2 Clinical perspective

Unlike many complex chronic and lifestyle-associated diseases, infectious diseases are usually caused by a single agent. In turn, identification of this agent typically points to disease-control measures (e.g., sanitation) as well as treatment (e.g., vaccination) [7] and tools to identify agents responsible for a presented infection have been widely sought. The traditional microbiology lab methods for detecting and identifying bacterial pathogens include Gram staining, liquid or solid culture, and the use of the live microbes to assay for antibiotic resistance [2].

Conventional laboratory methods exhibit a trade-off between resolution scope. Culture has favorable scope, meaning that many bacterial pathogens grow in culture and can be identified. However, not all bacterial pathogens can grow in culture. Culture is either not suitable or must be adapted for other pathogens, such as fungi or viruses. As a result, slow-growing, non-bacterial, or exotic pathogens can prove difficult to identify with culture. Furthermore, resolution may be poor, meaning that there may be - for example - no way to distinguish between strain or species with culture. On the other hand, DNA-based methods such as qPCR have high resolution. They typically can identify a single micro-organism at high (e.g., strain or species) resolution. Yet, the assay works only for a single micro-organism.

1.2.3 High-throughput methods

Some consider that the rapid identification of the SARS virus in 2003 ushered in a new era of pathogen identification. This was achieved through a combination of high-throughput techniques (nucleic acid microarray hybridization) and traditional viral culture and real-time PCR [2]. Since that time, high-throughput techniques, such as MALDI-TOF mass spectrometers, have become gradually introduced to clinical workflows. By comparing protein signature in a clinical sample with a collection of patterns that have been deposited in a database, MALDI-TOF can often achieve better resolution and faster turn-around time than culture [9]. Yet, the discriminatory power of the method varies depending on the target micro-organism as well as the database used. Some bacteria are under-represented in MALDI-TOF databases,

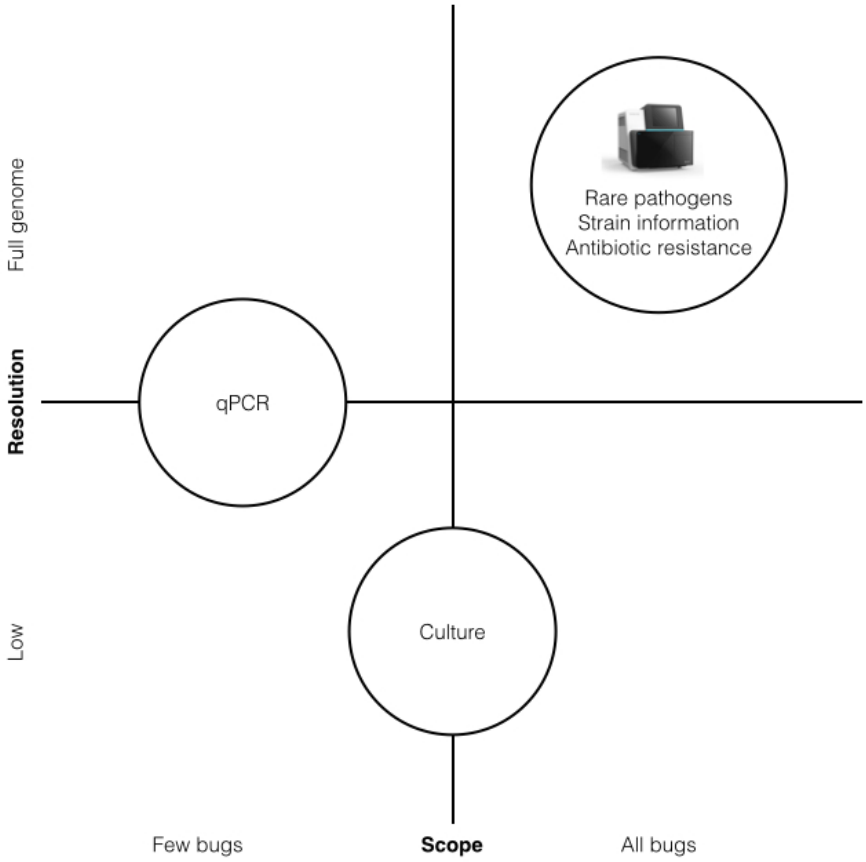

Figure 1.2: The trade-off between scope and resolution.

technical problems (e.g., variations in culture or sample preparation) can affect the discriminatory power, and many commercial databases do not include viruses.

Soon after the publication of the Sanger-sequenced human genome draft results in 2001 and the "finished" sequence in 2004, several new DNA sequencing technologies were described in the literature. Most used a flow-cell surface or beads in an emulsion to spatially segregate individual DNA template molecules so that they can be amplified *in situ* and sequenced in parallel with simultaneous data acquisition from millions of templates via optical or electronic detection [2].

These technologies ushered in an era of next-generation DNA sequencing (NGS). As costs drop and performance improves, NGS is becoming an appealing alternative (or supplement) to MALDI-TOF, culture, or targeted DNA-based methods like qPCR [17]. The critical advantage of NGS in for infectious diseases is that it can, in principle, assay every gene and every conceivable marker derived from infectious agents in a sample. Whereas MALDI-TOF relies on a handful of signature proteins, NGS is capable of identifying an unlimited set of possible pathogens (unlimited scope) as well as the complete genomic sequence of each one (high resolution) (Figure 1.2). With sufficiently long read lengths, multiple reads mapping to a specific microbial genomes, and a well-annotated reference database, nearly all microorganisms can be uniquely identified on the basis of their nucleic acid sequence [2].

1.2.4 The case for high-throughput sequencing

There are two central reasons driving interesting in unbiased NGS for comprehensive detection of pathogens from clinical samples: (1) Conventional diagnostic testing for pathogens still fails to detect the causal agent in a significant percentage of cases [12]. (2) Failure to accurately diagnose and treat infection in a timely fashion contributes to continued transmission and increased mortality in hospitalized patients. Furthermore, a rising tide of studies have collectively made a strong case for the introduction of NGS into the clinic for various compelling scenarios.

Unbiased screening of rare pathogens: Because NGS performs unbiased measurement of all nucleic acids in a clinical sample, it can reveal pathogens that escape

conventional clinical testing. A recent study applied NGS to a 14-year-old boy with severe immunodeficiency who presented with fever and headache that gradually progressed to hydrocephalus and status epilepticus [19]. Conventional diagnostic workup, including brain biopsy, was unrevealing. Yet, unbiased next-generation sequencing of the cerebrospinal fluid identified *Leptospira*, an exotic pathogenic bacteria. Though conventional clinical assays for leptospirosis were negative, detection with NGS informed intervention and allowed the patient to make a full recovery.

Outbreaks: Microbiologists and physicians often need to look broadly before determining the virulence genes in a particular strain account for an outbreak. NGS facilitates this search. For example, NGS was used to identify a novel strain of *Escherichia coli*, O157, during a food-borne outbreak in Germany [9]. Similarly, NGS has been applied to the US epidemic of community-associated methicillin-resistant Staphylococcus aureus (MRSA). NGS indicated that most strains were very closely related across geographical locales, implicating expansion from a single population rather than convergent evolution of different strains [2]. As a finally example, NGS applied to the Haitian cholera outbreaks traced its probable origin to UN solders that inadvertently brought the infection from Bangladesh [2].

Resistance: NGS cam be used to determine whether plasmids or other mobile genetic elements carrying antimicrobial drug-resistance genes are being transferred among the bacterial pathogens infecting patients. For example, the NIH recently experienced an outbreak of carbapenem-resistant *K. pneumoniae* that affected 18 patients and killed 11. Integrated genomic and epidemiological analysis traced the outbreak to three independent transmissions from a single patient who was discharged 3 weeks before the next case became clinically apparent and pointed to possible explanations for these transmissions [9]. Similarly, NGS applied to patients infected with HIV has been used to reveal viral subpopulations and low-frequency mutant viral strains with antiviral resistance?associated sequence changes [2].

Culture-free: NGS is valuable in clinical settings when dealing with difficult-to-culture or notoriously slow-growing pathogens such as *Mycobacterium tuberculosis*.

Microbial populations: NGS can be used to explore microbial diversity and full populations. Nine Mycobacterium species can cause tuberculosis. Some strains,

such as *Mycobacterium bovis*, require specific antibiotic treatments, making high resolution NGS particularly valuable [9]. Furthermore, the human microbiome project has highlight the importance of assaying microbial populations [4]. Community-wide profiling and examination of changes in relative abundance may become increasingly important as we learn more about human microbiology.

1.2.5 NGS challenges and opportunities

Cost and speed: The cost and turn-around time for sequencing have both been driven down by hardware advances. The cost for determining individual microbial genomes continue to fall and costs as little as $100 per sequence [9] with multi-hour turn-around. Both will continue to improve and the justification for NGS will become increasingly apparent, starting with hospital patients who develop difficult-to-treat or life-threatening infections that prove very costly to the system.

Informatics: NGS technology produces large datasets that require extensive bioinformatics simply for sequence analysis. Data presentation and distillation of clinical recommendations from large datasets also prove challenging. Addressing informatic challenges associated with NGS will be critical for widespread adoption.

Mechanism: Increasingly, NGS has been applied to the molecular networks that underlie cells, including chromatin immunoprecipitation with subsequent high-throughput sequence analysis (ChIP-Seq) for protein-DNA interactions, high-throughput RNA sequencing (RNA-Seq) for transcription, Ribo-Seq for translation, parallel analysis of RNA structure (PARS) for structure assays, and global mapping of DNA-DNA interactions using proximity ligation coupled with deep sequencing (Hi-C) [17]. Many of these methods could also be applied to the study infectious disease.

1.3 Contributions and outline of this thesis

1.3.1 Informatics: A new paradigm for NGS-based diagnostics

The first part of this thesis presents a new paradigm for infectious disease diagnostics. We have shown that NGS can be used to isolate and count micro-organism derived cell-free DNA fragments in human blood. We built a pipeline and application for processing and browsing this data. We applied this technique to thousands of samples and hundreds of patients at Stanford hospital, showing that this methods works for viral detection as well as deep tissue microbial infections. We further showed that unbiased screening via NGS can reveal rare or un-expected infections.

1.3.2 Mechanism: A new strategy for mechanistic studies

The second part of this thesis presents a new strategy for studying infectious disease mechanism. We have shown that iCLP-seq, an NGS-based methods for assaying genome-wide RNA-protein interactions, can be used to study the interaction networks between hosts cells and viruses. Because these networks are central for viral replication, the pipeline we developed may provide insight into disease mechanism.

Chapter 2

Infectome pipeline

2.1 Human microbiome in cell-free DNA

The human microbiome is now recognized as an important contributor to human health [4]. NGS applied to external body sites have been very fruitful, as these sites can be easily sampled in large cohort studies in order to assemble populate-wide statistics on microbial composition. Yet, little is known about the microbial composition of deeper tissues, as it is challenging to access them.

Parallel efforts have focused using NGS as a tool to assay blood. These efforts take advantage of a phenomenon first described in 1947 by Mandel and Metais [16]. They discovered that blood contains circulating cell-free DNA. These DNA fragments enter blood as the detritus of dead and apoptosed cells, and are likely nucleosome-protected fragments enriched in circulation. Methods of molecular counting, notably NGS, have taken advantage of this phenomenon, using a new era of universal noninvasive diagnostic tests. Starting with detection of aneuploidies (such as Down syndrome) for pregnant women [6], molecular coating by NGS has been extended to organ transplants [5] (which can be thought of as genome transplants) and cancers [13]. The in the latter two cases, specific mutations can be used to resolve donor- or cancer-derived DNA fragments in circulation.

DNA species may correlate with health, their greatest value has been to measure the pro- portion of foreign genomes within an individual. Cancer-related mutations

can be used to some extent to determine the progress of disease (2), male chromosomal markers can be used for prenatal sex determination (3), and, in a sim- ilar fashion, women who have received organ transplants from males can be monitored for the amount of organ- specific DNA.

In a recent study, it was shown that non-human (microbial, fungal, and viral) derived cell-free DNA fragments can purified from blood and counted using NGS [5]. While translocation of microorganisms, or microorganism components, from the lumen of the GI tract into the systemic circulation can occur, it often has detrimental consequences, including activation of the immune system, and in extreme cases it can lead to septic shock [3]. The absence of acute sepsis of similar symptoms in our cohort further suggests that the majority of detected organisms are innocuous detritus of dead cells. Nevertheless, their existence suggests that microorganism-derived cell-free DNA fragments in blood can serve as a snapshot of microbial composition in diverse body sites and / or provide a way to monitor infections.

2.2 Pipeline for capturing microbiome in cell-free DNA

The general strategy for read assignment for any clinical sampled has been well-described. Computational subtraction of reads corresponding to the host (e.g., human) is typically performed using short-read aligners (e.g, Bowtie) and is followed by alignment (e.g., BLAST) to reference databases that contain sequences from candidate pathogens (e.g., NCBI). Yet, four challenges frustrate the use of NGS for infectious disease detection in cell-free DNA. Published strategies for detection of microorganism derived cell-free DNA follows this general workflow [5] though also employs an algorithm to reduce ambiguity in these alignments [20]. In spite of this, four challenges underlie this analysis.

Large data Alignment and classification algorithms must contend with massive amounts of sequence data. Recent advances in NGS technologies have resulted in instruments that are capable of producing >100 gigabases (Gb) of reads in a day. Furthermore, reference databases of host and pathogen sequences used by BLAST range in size from 2 Gb for viruses to 3.1 Gb for the human genome to 42 Gb for all

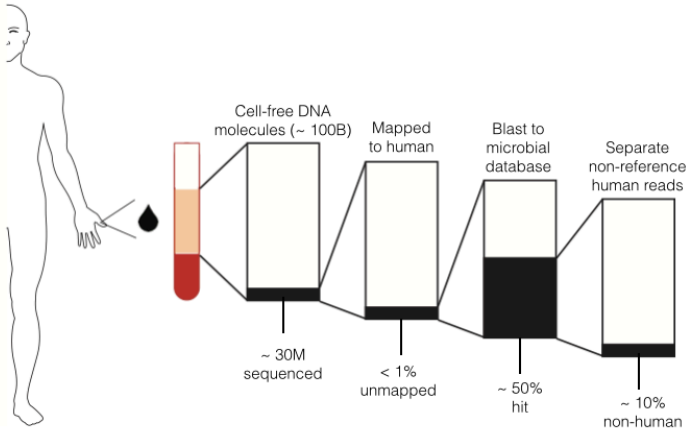

Figure 2.1: Isolation of non-human cell-free DNA.

nucleotide sequences in the National Center for Biotechnology Information (NCBI) nucleotide (nt) collection (as of January 2013).

Signal Only a small fraction of short NGS reads in clinical metagenomic data typically correspond to pathogens. This is particularly true in the case of cell-free DNA, as non-human cell-free DNA is $< 1\%$ of mapped reads. Furthermore, only a fraction of these un-mapped reads are actually derived from non-human sources, as most is either human DNA that is not found in the reference index used for mapping or does not align to the BLAST database used after mapping (Figure 2.1).

Speed BLAST is likely too slow for routine clinical analysis of NGS metagenomics data, as end-to-end processing times, even on multicore computational servers, can take several days to weeks. Analysis pipelines that use faster, albeit less sensitive, algorithms upfront for host computational subtraction, such as PathSeq, still rely on traditional BLAST approaches for final pathogen determination.

Interpretation the data must bed organized and presented at scale, across large clinical cohorts, such that it is intuitive for researchers and clinicians.

The signal problem will largely be addressed through bio-chemical methods to enrich for non-human derived nucleic acids. Furthermore, the speed problem has recently been addressed using faster alignment algorithms, such as SNAP in place of BLAST and RAPSearch for assignment of de novo contig assemblies in order to

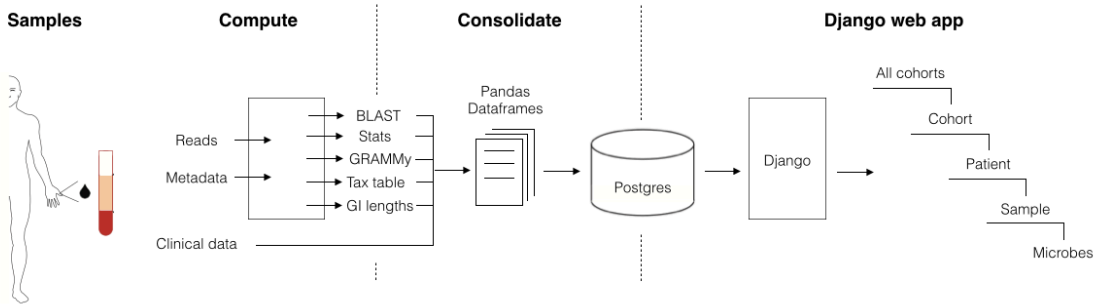

Figure 2.2: Django application for infectome data.

classify potentially novel organisms [12]. Despite these technical advances, there will probably continue to be a gap between the availability of such data and the ability to comprehensively interpret the results for clinical decision making. With this in mind, our emphasis was to develop a full application stack that performed both processing and visualizations to aid interpretation of all microorganisms in cell-free DNA sequence data.

In order to achieve this, we developed an application stack written in Python that used the Django web-development framework with the Matplotlib visualization library [10] and the Pandas library for data analysis. We built the Django application on top of a Postgres database using Sqlalchemy. The first part of the pipeline work as described previously, using several algorithms to compute relative abundance of microorganisms in each sample [5]. Reference files (e.g., taxonomic table), output files (e.g., BLAST results), and sample -meta data then written to Postgres. The tables in Postgres are referenced directly by the Django application (Figure 2.2).

Browsers have been broadly useful after their emergence in the early 2000s in response to rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data [14]. In turn, our application was designed to be data browser that can be used to explore cell-free DNA infection (or *Infectome*) data at all relevant scales. Indeed, we designed the application for intuitive navigation of this multi-scale data: each cohort is comprised of patients, which in turn may have many samples. In each sample, there may be thousands of unique

infections identified in the cell-free DNA sequencing. Furthermore, infections may be viewed at different levels of taxonomic complexity, such as genus or species.

With this in mind, we designed the Django application to present a series of web pages that reflect each relevant level of organization in the data. The cohort page reflects the highest level of organization. It presents a table of patients, which is sorted by the number of samples per patient, that provides a link to explore data for each patient in the cohort. It also provides cohort-level histograms that explain both incidence of the identified infection (fraction of samples in which each identified infection is found) as well as the load per sample (the number of infections identified per sample). In addition, it provides a table of sorted infections by prevalence within the cohort. Finally, it provides a toggle that allows the data to be presented at different levels of taxonomic resolution, from genus to species (Figure 2.3). Collectively, this make is possible to navigate the data in two primary ways: it is possible to take a patient-centric approach to the data and examine data for specified patients. Or, it is possible to take an infection-centric approach to the data and examine simply the infections identified in the cohort across patients.

The patient-centric approach can be used to quickly identify the infections identified within a specified patient at a specified taxonomic scale (e.g., genus or species). In order to present this information intuitively, we transform the raw abundance measurements returned by the sequencing pipeline The pipeline uses an algorithm (GRAMMy) to process the raw BLAST results; GRAMMy addresses two problems. First, each organism has a different genome size and, in turn, genome size affects the number of reads expected for each. Second, reads often align to multiple genomes. Taking these into account, GRAMMy performs a maximum likelihood estimation for read assignment to each organism and provides relative abundance measurement per organism per sample. From this measurment, we back out an estimate for absolute read counts per genome. With this value, we compute a coverage ratio between each infection and human for that sample and scale this value by 10^6 to get relative genome copies per million (gcm). In isolation, this value is not particularly intuitive: it indicates the number of genome copies for a given infection relative to human in that sample. For presentation, we simply compute a percentile

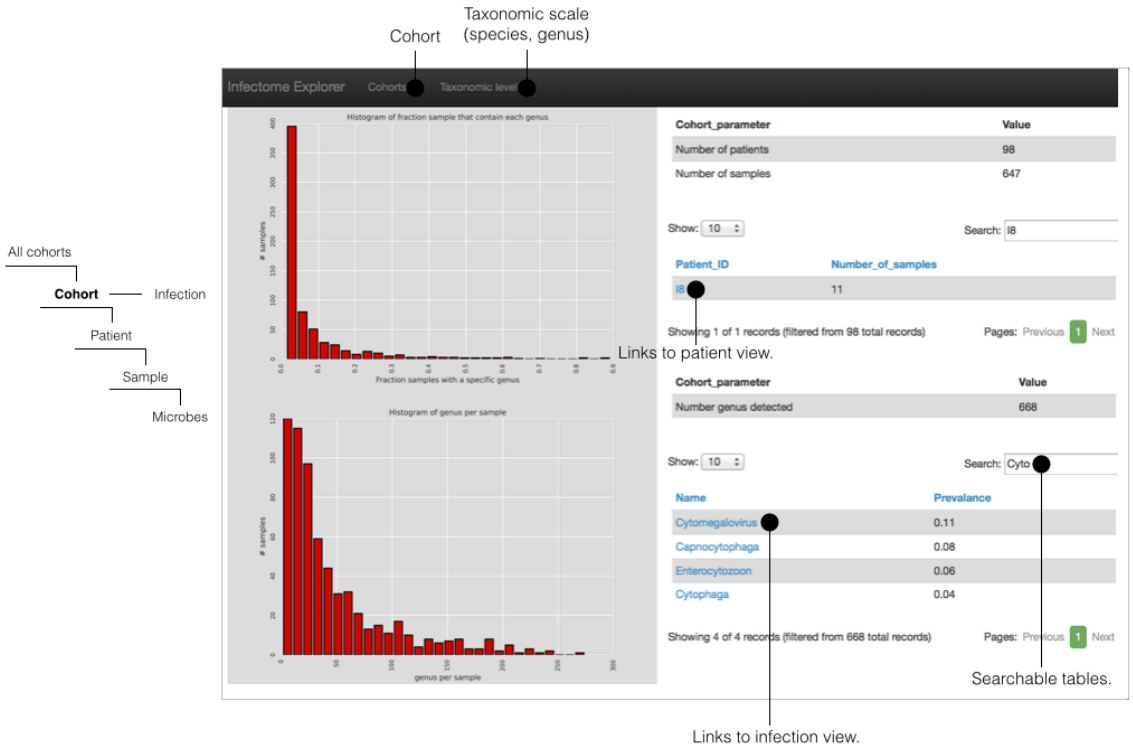

Figure 2.3: Cohort data

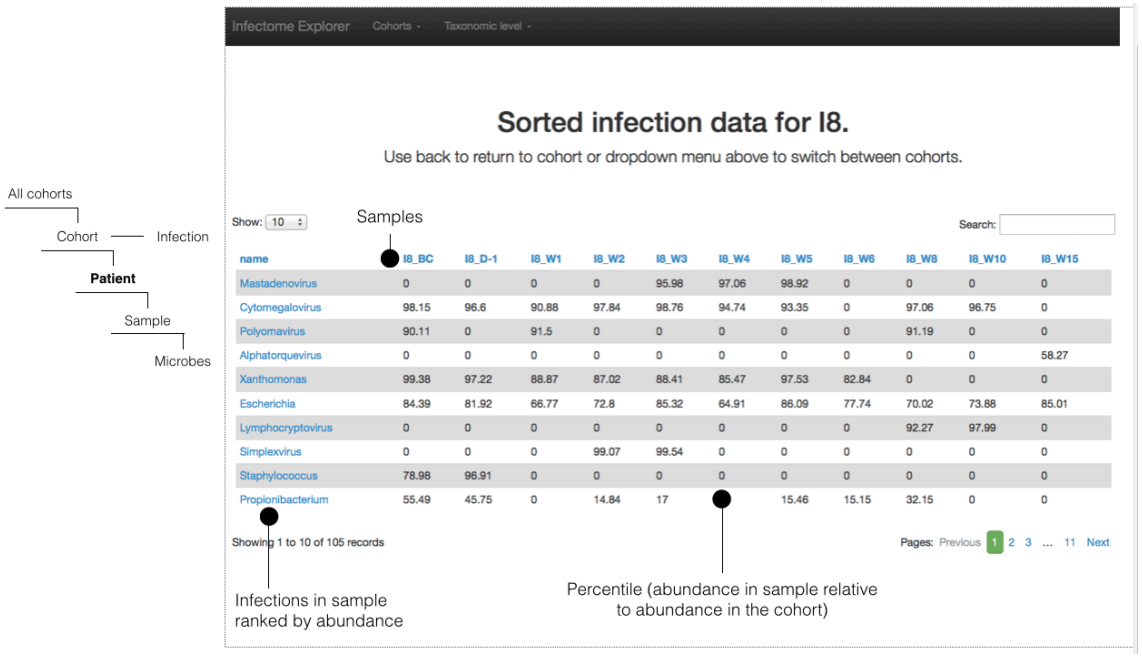

Figure 2.4: Patient data

(Figure 2.4) for each gcm value with respect to the cohort for that infection. In turn, the percentile indicates the magnitude of each measurement relative to what is observed across the cohort.

From the patient view, it is possible to drill down into each identified infection. In this case, it is useful to know both the timeseries data for that infection as well as more detailed information about read coverage across the microbial genome. Both measurements can provide greater confidence about the legitimacy of a given signal. For example, a consistent infection timeseries across samples supports likelihood of a bona-fire infection relative to a spurious signal found in one sample or a higher irregular pattern. Furthermore, coverage is computed simply by aggregating GIs (individual sequence records in the BLAST database) for each taxID, lining the GIs up continuously into a composite "genome", and then evaluating reads with respect to position across this composite genome. Irregular coverage patters may be indicative of database contamination whereas consistent patters across the composite genome support presence of the identified infection. Using data from a bone

Figure 2.5: Infectione timeseries and coverage

marrow transplant cohort, we show both timeseries and coverage data collected for a particular patient (I8), Figure 2.5).

To demonstrate clinical use of this application, we consider the case of I6, a pediatric bone marrow patient with severe respiratory complications. We collected longitudinal cell-free DNA samples across over the course of post-transplant therapy and processed the data with our application. From the application views, it was immediately clear that I6 a had very load of a rare Polyomavirus species, WU polyomavirus, that has been implicated in severe respiratory illnesses Figure 2.6). Though the patient was tested for a similar polyomavirus, BK virus, those tests were negative. Indeed, the situation is similar to a scenario recently described in which NGS of cerebrospinal fluid identified an exotic pathogenic bacteria, leptospira, that explained severe hydrocephalus and status epilepticus [19]. In this case, the patient died prior to clinical intervention based upon this information. However, the case does highlight the fact that unbiased and broad screening of potential pathogens in

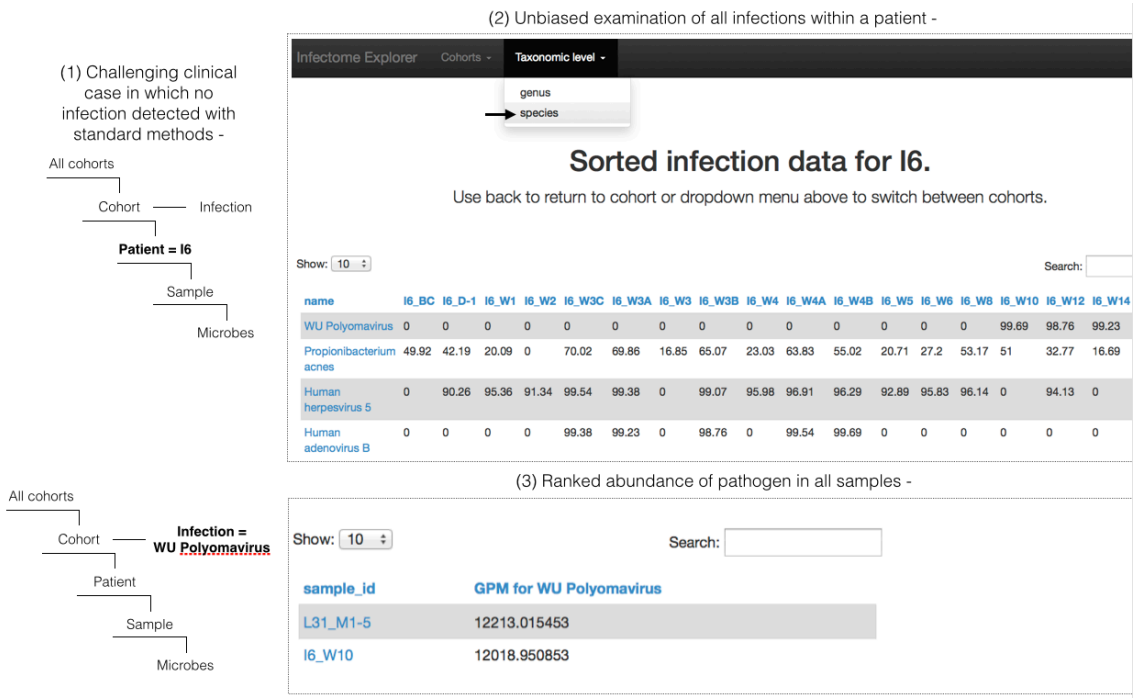

Figure 2.6: Clinical use of infectome application

severe cases can reveal agents that escape conventional clinical testing and, in turn, will likely be a powerful supplement to existing clinical assays going forward.

Chapter 3

Clinical validation

3.1 Organ transplantation

In organ transplant recipients, the ratio of recipient genomic DNA to graft-derived donor DNA, distinguished by SNPs that are specific to the recipient or the donor, provides a measure of the number of graft cells that are dying and releasing their DNA into the blood. In a pilot study of heart transplant recipients, episodes of acute cellular organ rejection were marked by increases in the proportion of donor-derived DNA in the blood. Advantages of this approach over traditional periodic biopsies of the graft tissue are that it is less invasive, may be less affected by sampling errors if lymphocytic infiltrates or regions of cell [18].

We first applied the infectome pipeline to existing cell-free DNA samples banked and processed by the Quake lab. These samples were largely from organ transplant recipients because rejection can be monitored by quantifying cell-free donor-specific DNA in the transplant recipient?s plasma via shotgun sequencing (?genome transplant dynamics?, GTD). By using single nucleotide polymorphisms (SNP) to discriminate between donor and recipient DNA molecules, GTD provides a non-invasive yet direct measure of graft damage. In this cohort, immunosuppressive therapies significantly reduce the risk of graft rejection, but increase the susceptibility of recipients to infections. Together with allograft rejection, infectious complications remain one of the most important causes of morbidity and mortality after

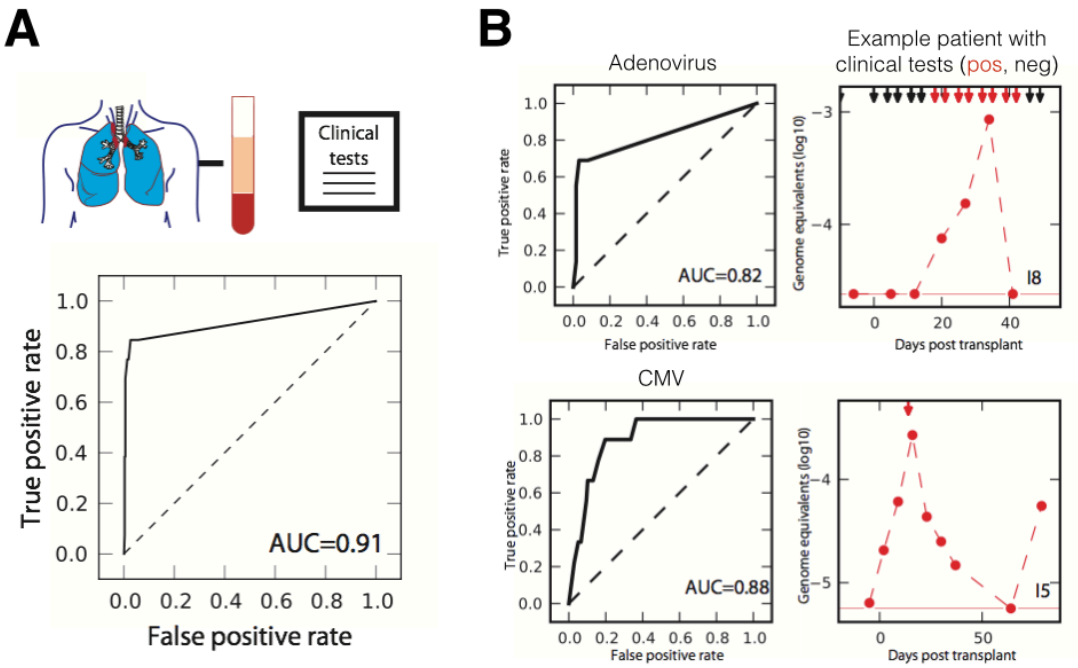

Figure 3.1: Clinical correlations on viruses

lung transplantation, with cytomegalovirus infections (CMV) posing the most significant known threat.

With this in mind, we explored whether cfdDNA levels correlate with clinical indicators of infection. Infectious pathogens are identified by simultaneously identifying non-human cfDNA sequences and comparing them to known genomic databases of bacterial, viral and fungal pathogens. To study the relationship between infection and graft damage, we collected over 35000 clinical measurements of specific infections performed on 14 specimen types for the lung transplant cohort. We first quantified reads that map to the CMV genome for each sample and observed increased CMV abundance in samples that were clinically positive for infection, resulting in an AUC of 0.91 Figure 3.1). This data indicates that CMV surveillance can be performed in parallel with rejection monitoring using the same sequence data and led us to examine whether other viral infections could be similarly monitored.

We next performed a similar analysis on a pediatric bone marrow cohort. Like

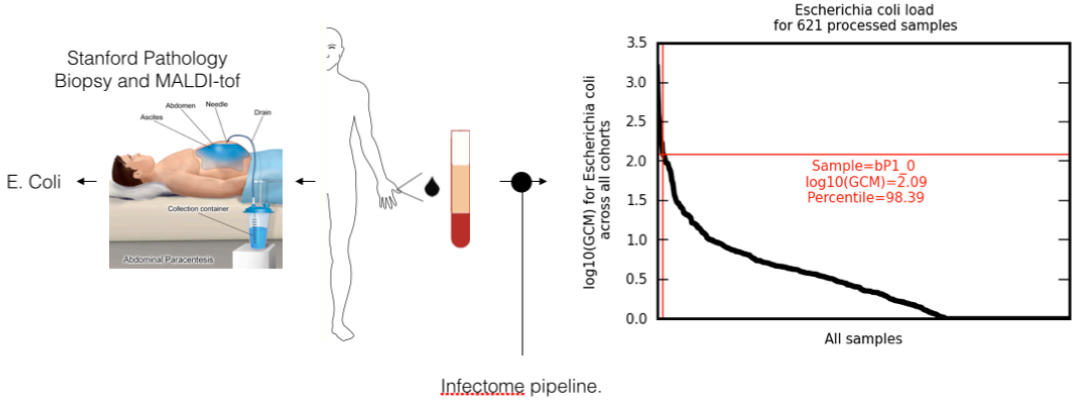

Figure 3.2: Clinical correlations with deep tissue sampling

lung, clinical testing for CMV was common. Because of the elevated risk in the pediatric cases, this cohort includes regular screens for additional viruses. For example, Adenovirus is a community-acquired respiratory infection that can cause graft loss in transplant recipients and poses a particularly high risk for paediatric patients. The correlation between infection-derived cfDNA and positive clinical tests for viruses are evident in timeseries data, at elevation in signal is observed when positive clinical test results are recorded (Figure 3.1), and is further captured at cohort-level with ROC curves that have favorable AUC values of 0.82 and 0.88 for Adenovirus and CMV, respectively.

Collectively, the favorable performance on viruses is reasonable. Existing clinical test are performed on blood using molecular diagnostics, such as qPCR. With this in mind, the observed performance of NGS on cell-free DNA is not surprising. However, it is worth noting that the performance is quite encouraging considering the fact that these samples are not enriched for non-human signal. Indeed, infection-derived cell-free DNA is sparse in the sequencing libraries and, therefore, the favorable performance should only be expected to improve as greater depth is reached through enrichment strategies.

3.2 Deep tissues

We next examined the ability to resolve deep-tissue infections using blood as the sampled medium. This would be particularly appealing, because invasive biopsies to test for infection can be problematic for patients with compromised heath and can introduce risk. To examine this, we compared shotgun sequencing of blood biopsy sampled from a gastrointestinal abcess tested using MALDI-tof. The biopsy results (positive for E. Coli) agree with the cfDNA sampling, which shows that this patient have a very higher percentile E. coli measurement relative all all samples processed (Figure 3.2).

3.3 Untested infections

Considering the favorable results on both vises as well as bacteria, we further examined the merits of hypothesis-free screening by examine data for the lung transplant cohort. We identified well characterized pathogenic and onco-viruses as well as commensal torque teno viruses (TTVs, alphatorquevirus genus) in that data, is consistent with previous observations of a link between immunosuppression and TTV abundance. The frequency of clinical testing for these viruses varied considerably, with frequent surveillance of CMV (Human Herpes Virus 5, HHV-5) relative to other pathogens (Figure 3.3).

We evaluated the incidence of infection (number of samples in which a given virus is detected via sequencing) relative to the clinical screening frequency. Although CMV was screened for most frequently (335 samples), its incidence as determined by sequencing (detected in 22 samples) was similar to that of other pathogens that were not routinely screened, including adenovirus and polyomavirus (clinically tested on four occasions and one occasion, respectively). We further showed that hypothesis-free infection monitoring revealed numerous un-tested pathogens, including un-diagnosed cases of adenovirus, polyomavius, HHV-8, and microsporidia in patients who had similar microbial cfDNA levels compared to patients with positive clinical test results and associated symptoms.

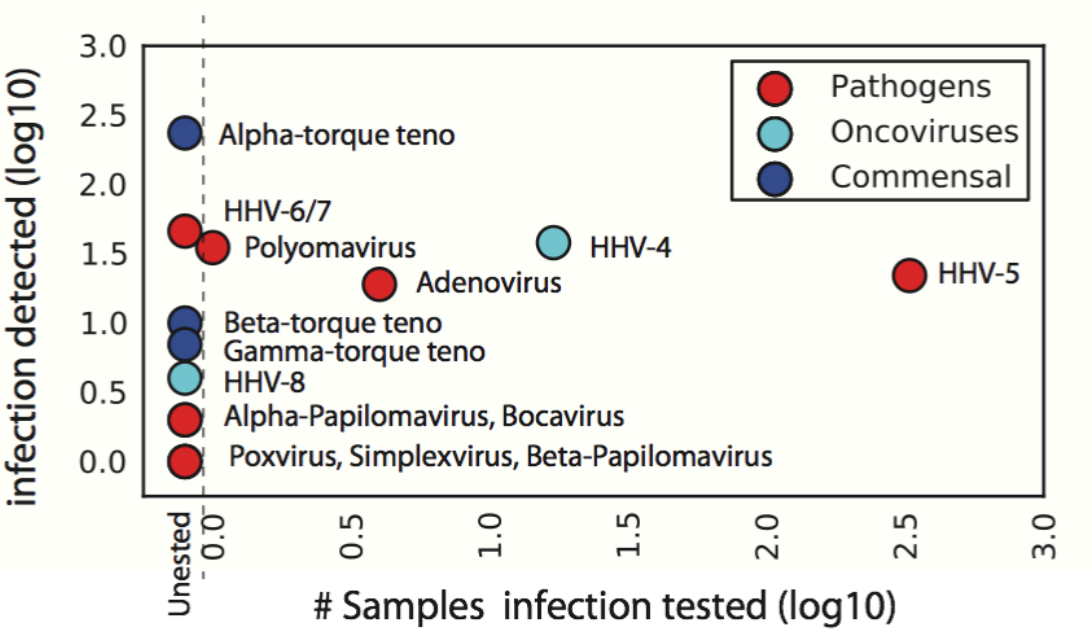

Figure 3.3: Clinical correlations on viruses

Chapter 4

The cell-free microbiome

4.1 Importance of the microbiome

In an adult human, the bacteria within the colon outnumber host cell numbers by up to two orders of magnitude. This results in a frequency of bacterial genes at least 100 times greater compared with those within the human genome [3]. The importance of the microbiome is evident in studies of germ-free animals: they are more susceptible to infections and have reduced vascularity, digestive enzyme activity, muscle wall thickness. The microbiota also leads to enhanced integrity of the structural barrier of the GI tract by metabolizing dietary carbohydrates into short-chain fatty acids, metabolizes toxic and potentially carcinogenic compounds such as pyrolysates, and produces biotin, folate, and vitamin K from dietary precursors, which are then absorbed by the GI tract and circulated. In turn, numerous studies have implicated an altered balance in the composition of the microbiota (dysbiosis) in many diseases, such as obesity, celiac disease, type 2 diabetes, atopic eczema, asthma, inflammatory bowel disease (IBD), and chronic diarrhea [3].

4.2 Linking blood and body sites

The Human Microbiome Project first defined the compositional range of the normal microbiome of healthy individuals [4]. Since this pioneering work, studies of the microbiome in different physiological contexts have been performed. Pregnancy is one important example and work has been driven by intense interest in the preterm birth problem, one of the leading causes of neonatal mortality in the United States. Until recently, the paradigm was that the majority of intrauterine infections originated in the lower genital tract with microbiota ascending into an otherwise sterile environment resulting in infection of the placenta and fetus [15]. It is now known that the microbiome changes during pregnancy, driven by hormonal and physical fluctuations [11]. Furthermore, recent work has shown that the human placenta is not sterile. Rather, it harbors a unique microbiome with a compositional signature close to that of the oral cavity. In turn, the placenta may be seeded by microbes traveling through blood from other body sites and changes in the placenta microbiome may influence adverse pregnancy outcomes, such as pre-term birth or infections [1].

Parallel work has shown that non-human (microbial, fungal, and viral) derived cell-free DNA fragments can purified from blood and counted with next-generation sequencing (NGS) [5]. Must of this material is likely to be the detritus of dead and apoptosed cells, with genomes chewed up and assayed via NGS like any other blood analyte [16]. Yet, the source of this material is unclear; few studies have even shown that non-human derived cell-free DNA fragments can be assayed and there have been no studies linking this material to body sites of origin. With this in mind, we assayed both blood and several body sites for sixty samples collected from a cohort of pregnant women at Stanford hospital.

For each blood sample, we isolated and cataloged microbial-derived DNA sequences using a metagenomic pipeline and further extracted all bacteria-derived reads for each sample. For these same samples, we obtained temporally matched 16s sequencing data from four body sites (Saliva, Vagina, Gut, and Gum). We first

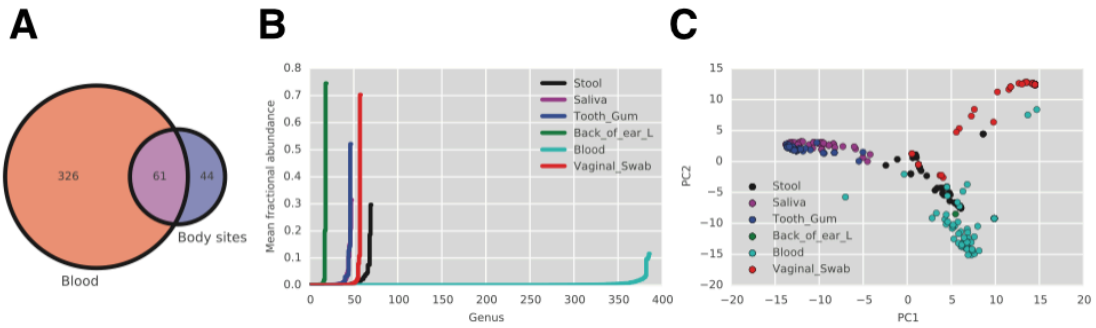

Figure 4.1: Composition of the blood microbiome

performed descriptive analysis of the data by comparing the taxonomic composition of blood relative to the sampled body sites. We discretized the abundance data for blood and all sampled body sites, resulting in a binary value for each genus in all samples. We then compared the genus detected in blood to the aggregated genus detected in all body sites. $\approx 58\%$ of the genus detected in any body site were also found in blood, while only $\approx 15\%$ of the genus detected in blood were also found in any body site. This suggested that micro-orgaism derived material in blood originates from more than just the sampled sources (Figure 4.1).

We then computed mean fractional abundances for each site at genus-level resolution. The 16s sampled body sites showed strong enrichment in particular genus that are known to be well-adapted to each biological niche [4]. Blood is quite different: the number of genus detected is ≈ 8-fold greater than the body sites, with a mean fraction abundance ≈ 10-fold lower than the body sites. We then transformed the data using PCA in order to determine the genus that most strongly drive the measured variation in blood as well as the sampled tissues (Figure 4.1). PCA indicates that blood samples generally cluster together in taxonomic space and occupy a distinct composition relative to the sampled body sites. We examined genus that strongly contribute to the principal components in order to understand what distinguishes blood from the body sites: as expected, genus - notably, Acidovorax and Cupriavidus - that drive the variation are found at high fractional abundance in the blood sampled, but are nearly absent from the sampled body sites.

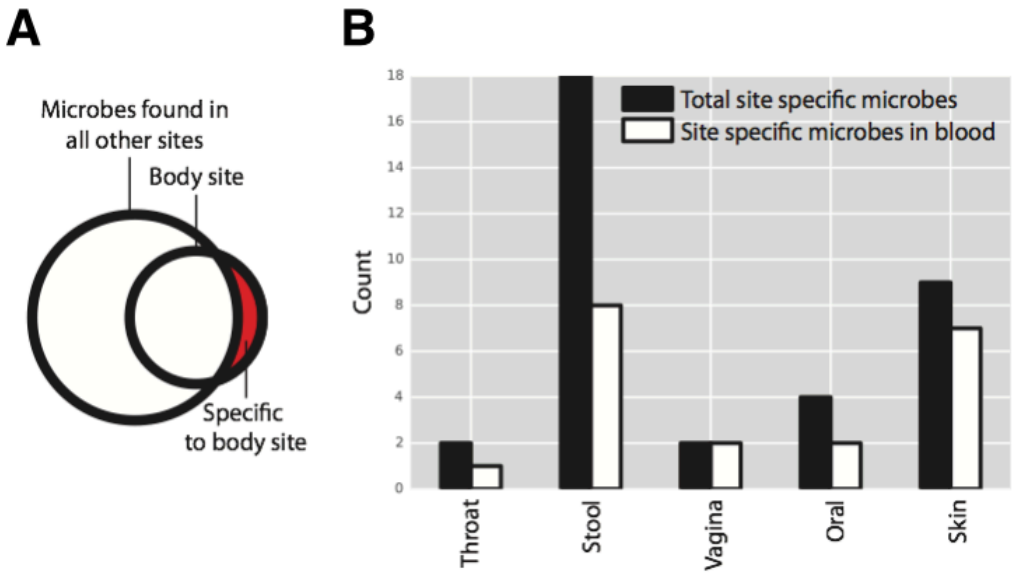

Figure 4.2: Detection of body site specific bacteria in blood

We next investigated whether blood samples microorganisms from each site. We reasoned that body-site specific micro-organisms provide a reasonable indicator for this. We compute specificity simply by discretizing the genus found in each body site and comparing there to genus found in all other sites (Figure 4.2). To aid this analysis, we used our sampled body site data as well as the metagenomic community profiles made available by the Human Microbiome Project [4], which contains 35 billion reads taken from 690 samples from 300 US subjects, across 15 body site. In both cases, we computed a list of body-site specific genus, which are only detected in single body sites for all sampled collected. We then asked whether these genus were detected in blood: we detected 57% and 45% of the site-specific genus in blood using site-specific genus determined via HMP and this study, respectively. We investigated the body site abundance of specific genus found in blood relative to those absent from blood. We found no significant difference between the abundance in either partition, which argues against the possibility that all site-specific genus were found in blood, but under-sampling frustrated their detection.

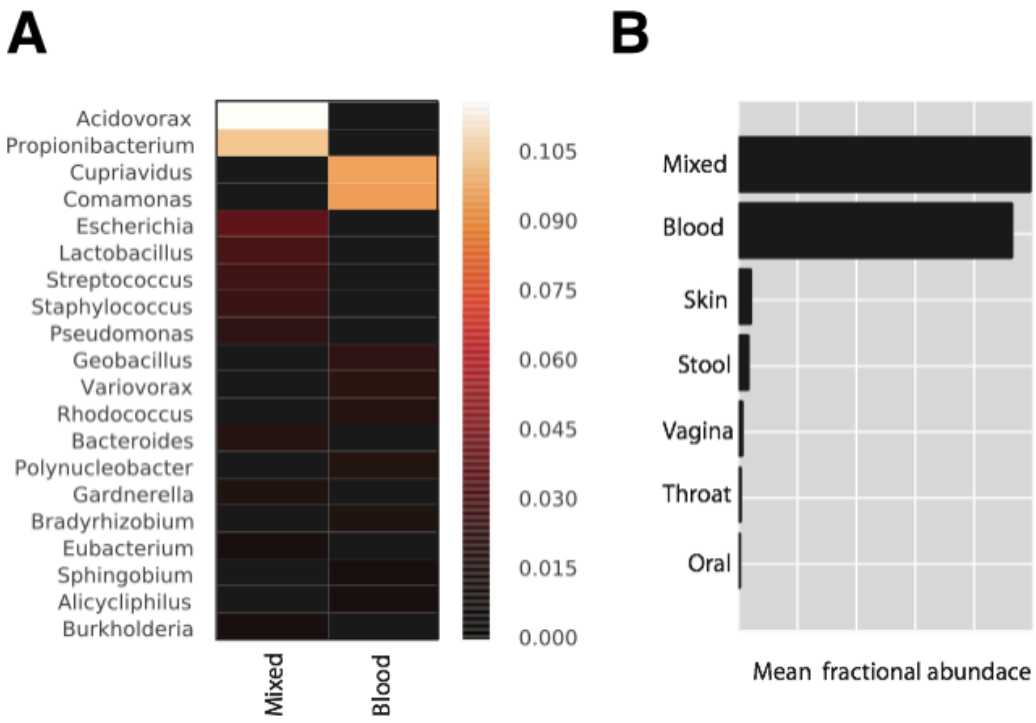

Figure 4.3:

From this analysis, we obtained an assignment of body site specific for each genus detected in our data. In turn, any genus could be assigned to a specific body site or mixed, meaning that it was detected in more than one tissue. We partitioned the genus detected in blood using this assignment in order to understand the origin of genus found in blood. These indicate that blood is composed primarily of two type: genus derived from mixed sources, which cannot be traced to any specific tissue, and genus that are specific to blood only (Figure 4.3).

4.3 Coupling between blood and body sites

We then asked if a physical relationship between blood and the body sites can be established based upon the data. Specifically, we examine whether the fractional abundance of genus at each body sites is linked to detection in blood. We discretized

the blood data for each sample. For each genus, we then evaluated the fractional abundance of that genus for all matched body site samples. We then compared the distribution of abundances at each body site based upon whether the genus was found in blood, expecting a difference (e.g., an elevation in abundance at a tissue) if there was coupling to blood (e.g., when the genus was found in blood). We found no signifiant evidence of coupling for all genus - body site combinations.

We further examined whether the composition of blood can be modeled as a function of the sampled sites. We chose a linear model, meaning that each blood sample is modeled as a linear combination of the genus at temporally matched body sites. We used a quadratic programming package to determine for mixing coefficients that minimize the squared error between the model and the blood measurement, subject to intuitive constraints (e.g., the coefficients must be greater than zero). After confirming the model performed correctly on mock data, we applied it to all samples. We found that the model performed poorly, using coefficient of determination (R^2) between guess and actually blood data as a measure of performance. We examined residuals to understand the failure, and found - intuitively - that the model fails because many highly abundant genus in blood are not found in the sampled body sites. In order words, the sampled sources are insufficient to describe the composition of blood, as blood apparently sampled from additional sources or serves as an environment that amplifies specific genus non-linearly.

4.4 Summary

The distribution of fractional abundance for genus detected in blood has a different shape than the sampled body sites: blood contains many more genus, but at far lower fractional abundance, with a long tail of genus found at trace abundance. In contrast, sampled body sites are dominated by few genus at high abundance and a relatively short tail of low-abudnance genus. This reflects the fact that some genus are well-adapted to each body site niche. In turn, blood may serve as a common sink into which all tissues contribute dead cells, resulting in a passive environment of mixed DNA fragments that we sampled. We also found that most abundant genus

in blood are essentially absent from sampled body sites. This suggests that either that blood can samples from more sources (e.g., internal tissues) or that blood may be a niche for a certain, narrow sub-set of genus.

Analysis of site-specific genus provides evidence that blood samples each body site, as around half of site specific genus are found in blood when analyzed at both genus and species-level resolution using both our body site data as well as the HMP data. However, we did not find clear evidence of coupling between body sites and blood: there is not apparent difference in genus abundance at any body site based upon whether that genus is detected in blood in a temporally matched sample. This is not surprising: the blood is under-sampled, meaning that few of the present microorgaism-dervied fragments are actually sequenced in our data, and blood is a complex mixture represent many tissue sources. The latter point also frustrated efforts to model blood as a function of the sampled sites: blood data contains genus that were not detected in any of the sampled tissues.

Bibliography

[1] K. Aagaard, J. Ma, K. M. Antony, R. Ganu, J. Petrosino, and J. Versalovic. The placenta harbors a unique microbiome. *Science Translational Medicine*, 6(237):237ra65–237ra65, 2014.

[2] S. D. Boyd. Diagnostic Applications of High-Throughput DNA Sequencing. *Annual Review of Pathology: Mechanisms of Disease*, 8(1):381–410, Jan. 2013.

[3] J. M. Brenchley and D. C. Douek. Microbial Translocation Across the GI Tract *. *Annual Review of Immunology*, 30(1):149–173, Apr. 2012.

[4] T. H. M. P. Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[5] I. De Vlaminck, K. K. Khush, C. Strehl, B. Kohli, H. Luikart, N. F. Neff, J. Okamoto, T. M. Snyder, D. N. Cornfield, M. R. Nicolls, D. Weill, D. Bernstein, H. A. Valantine, and S. R. Quake. Temporal Response of the Human Virome to Immunosuppression and Antiviral Therapy. *Cell*, 155(5):1178–1187, Nov. 2013.

[6] H. C. Fan, Y. J. Blumenfeld, U. Chitkara, L. Hudgins, and S. R. Quake. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proceedings of the National Academy of Sciences*, 105(42):16266–16271, 2008.

[7] A. S. Fauci and D. M. Morens. The perpetual challenge of infectious diseases. *The New England journal of medicine*, 366(5):454–461, 2012.

[8] P.-E. Fournier, M. Drancourt, P. Colson, J.-M. Rolain, B. La Scola, and D. Raoult. Modern clinical microbiology: new challenges and solutions. *Nature Reviews Microbiology*, 11(8):574–585, Aug. 2013.

[9] J. L. Fox. Technology comes to typing. *Nature Biotechnology*, 32(11):1081–1084, Nov. 2014.

[10] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):0090–0095, 2007.

[11] O. Koren, J. K. Goodrich, T. C. Cullender, A. Spor, K. Laitinen, H. Kling Bäckhed, A. González, J. J. Werner, L. T. Angenent, R. Knight, F. Bäckhed, E. Isolauri, S. Salminen, and R. E. Ley. Host Remodeling of the Gut Microbiome and Metabolic Changes during Pregnancy. *Cell*, 150(3):470–480, Aug. 2012.

[12] S. N. Naccache, S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, J. Bouquet, A. L. Greninger, K. C. Luk, B. Enge, D. A. Wadford, S. L. Messenger, G. L. Genrich, K. Pellegrino, G. Grard, E. Leroy, B. S. Schneider, J. N. Fair, M. A. Martinez, P. Isa, J. A. Crump, J. L. DeRisi, T. Sittler, J. Hackett, S. Miller, and C. Y. Chiu. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, 24(7):1180–1192, July 2014.

[13] A. M. Newman, S. V. Bratman, J. To, J. F. Wynne, N. C. W. Eclov, L. A. Modlin, C. L. Liu, J. W. Neal, H. A. Wakelee, R. E. Merritt, J. B. Shrager, B. W. Loo, A. A. Alizadeh, and M. Diehn. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine*, pages 1–9, Apr. 2014.

[14] S. I. O'Donoghue, A.-C. Gavin, N. Gehlenborg, D. S. Goodsell, J.-K. Hériché, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, T. Walter, and B. Wong. Visualizing biological data—now and in the future. *Nature Methods*, 7(3s):S2–S4, Mar. 2010.

[15] A. L. Prince, K. M. Antony, D. M. Chu, and K. M. Aagaard. The microbiome, parturition, and timing of birth: more questions than answers. *Journal of Reproductive Immunology*, 104-105:12–19, Oct. 2014.

[16] S. Quake. Sizing Up Cell-Free DNA. *Clinical Chemistry*, 58(3):489–490, Feb. 2012.

[17] J. Shendure and E. L. Aiden. The expanding scope of DNA sequencing. *Nature Publishing Group*, 30(11):1084–1094, Nov. 2012.

[18] T. M. Snyder, K. K. Khush, H. A. Valantine, and S. R. Quake. Universal noninvasive detection of solid organ transplant rejection. *Proceedings of the National Academy of Sciences*, page 201013924, 2011.

[19] M. R. Wilson, S. N. Naccache, E. Samayoa, M. Biagtan, H. Bashir, G. Yu, S. M. Salamat, S. Somasekar, S. Federman, S. Miller, R. Sokolic, E. Garabedian, F. Candotti, R. H. Buckley, K. D. Reed, T. L. Meyer, C. M. Seroogy, R. Galloway, S. L. Henderson, J. E. Gern, J. L. DeRisi, and C. Y. Chiu. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *New England Journal of Medicine*, 370(25):2408–2417, June 2014.

[20] L. C. Xia, J. A. Cram, T. Chen, J. A. Fuhrman, and F. Sun. Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads. *PloS one*, 6(12):e27992, Dec. 2011.