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 fates of nations. Only two centuries ago, infectious diseases were the 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 effective treatments for infectious diseases. Washinton suffered from smallpox and multiple debilitating bouts of malaria, suffered wound infections and abscesses, nursed his brother on a tropical island as he died of tuberculosis, and had an influenza pandemic named after him [4]. Indeed, almost all the major advances in understanding and controlling infectious diseases have occurred in the past two centuries since the founding of the United States.

The first animal-transmission studies were conducted soon after the War of 1812 and were followed by the development of microscopes, which linked microorganisms to skin and mucosal diseases. Koch developed a unifying principle of infectious diseases in the late 1800s, proving four criteria designed to establish a causative relationship between a microbe and a disease. In the early 20th century Paul Ehrlich developed anti-infective serums and chemicals to destroy pathogens, which paved the way for vaccines, antibiotics, and antiviral agents that have saved hundreds of millions of lives and greatly extended the human life span.

1.2 Technology

1.2.1 Historical

Since the seminal contrabutions of scientists like Koch and Ehrlich, technology has driven our understanding of infectious disease. In 1980, only 1800 validated bacterial species had been published. A cornerstone of microbiology since the nineteenth century, cultivation *in vitro* fails for many microbes that have adapted to niches that are not easily recalculated in laboratory culture. With this in mind, the advent of DNA-based analyses enable direct identification of micro-organisms by directly reading regions of DNA or RNA genome and performing taxonomy based upon this digital signature. Hantavirus pulmonary syndrome, an ancient disease caused by a phlebovirus, was discovered unexpectedly in 1993 by the application polymerase chain reaction (PCR). Less than a year later, PCR-related subtraction techniques solved a century-old mystery of the cause of Kaposiś sarcoma, human herpesvirus 8. Since that time, DNA-based analyses have become far cheaper and are now accelerating the discovery of microorganisms (Figure 1.1) [5].

Unlike many chronic and lifestyle-associated diseases resulting from multiple, interacting risk cofactors, most infectious diseases are caused by a single agent. Identification typically points the way not only to general disease-control measures (e.g., sanitation, chemical disinfection, hand washing, or vector control) but also to specific treatment (e.g., vaccination or antimicrobial treatment) [4]. In turn, tools to rapidly identify the causative agent for a presented infection have been widely sought and developed. The challenge has historically been a trade-off between resolution scope. Whereas laboratory culture methods have reasonable scope (e.g., many organisms can grow and be identified via culture-based methods), the resolution is often poor. It often provides no way to distinguish between strain or species. Furthermore, many microorganisms do not effective grow in culture conditions. On the other hand, DNA-based methods such as qPCR have high resolution (e.g., very specifically identify a single microorganism) but only have a single target.

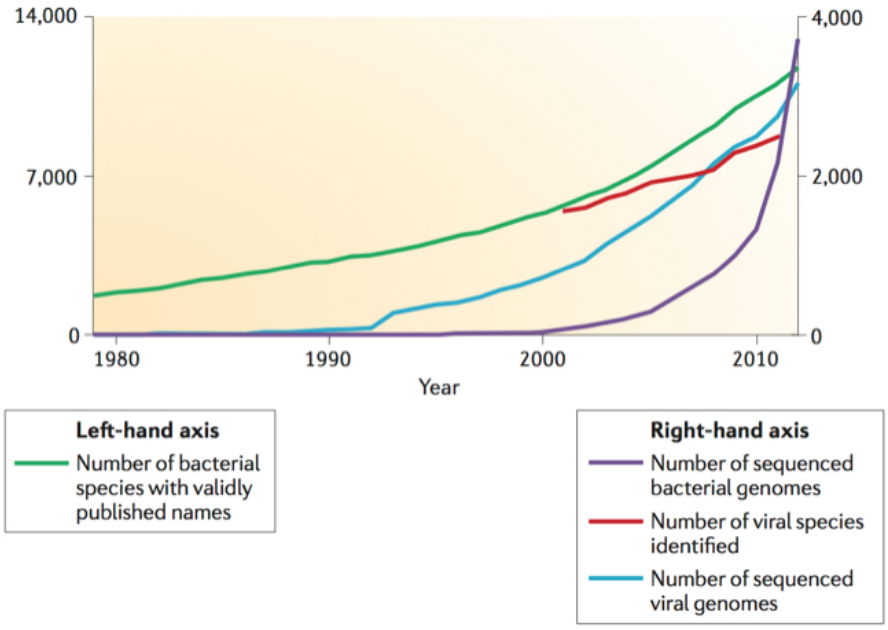

Figure 1.1: Rapid growth in the identification of micro-organisms.

1.2.2 High-throughput methods

MALDI-TOF mass spectrometers have become commercialized in order to better address the trade-off between resolution and scope. By identifying proteins with specific signature patterns in a clinical sample and comparing these signatures to a collection of patterns that have been deposited in a database, MALDI-TOF can achieve better resolution and faster turn-around time than culture-based methods [6]. Yet, the discriminatory power of the method varies depending on the species and the exhaustiveness of the database used. Some bacterial taxa are under-represented in MALDI-TOF databases, technical problems (such as variations in culture conditions, sample preparation and the spectrometer used) can affect the discriminatory power, and many commercially available databases do not include viruses.

With these challenges in mind, next-generation DNA sequencing (NGS) has become a very appealing clinical alternative or supplement to MALDI-TOF, culture, or targeted DNA-based methods like qPCR. There are numerous NGS platforms,

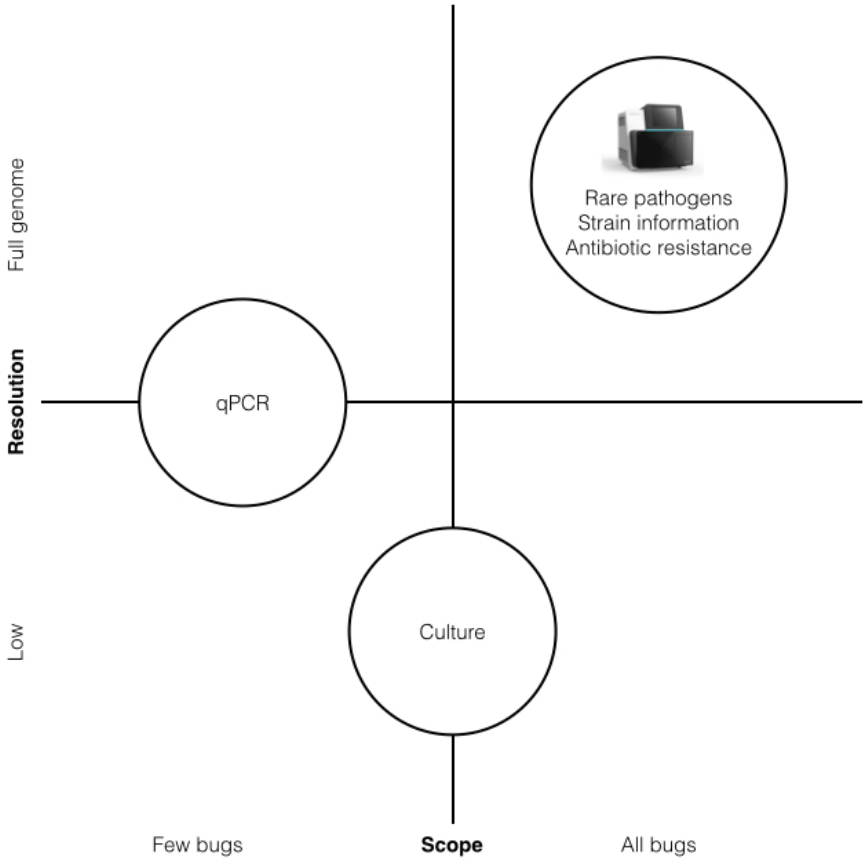

Figure 1.2: The trade-off between scope and resolution.

though all read a large population of amplified DNA fragments in parallel, producing thousands or millions of sequences concurrently [11]. 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 MS 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).

1.2.3 The case for high-throughput sequencing

Unbiased screening of rare pathogens: Because NGS samples all nucleic acids in a clinical sample in an unbiased fashion, it can reveal infectious causes of illness that escape conventional clinical testing. A recent study applied NGS to a 14-year-old boy with severe combined immunodeficiency that presented with fever and headache that progressed to hydrocephalus and status epilepticus. Though diagnostic workup including brain biopsy were unrevealing, unbiased next-generation sequencing of the cerebrospinal fluid identified an exotic pathogenic bacteria, leptospira. Conventional clinical assays for leptospirosis were negative, though detection with NGS informed intervention and the patient to make a full recovery [12].

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 foodborne outbreak in Germany [6].

Resistance: Whole genome sequencing can being used to determine whether plasmids or other mobile genetic elements carrying antimicrobial drug-resistance genes are being transferred among the bacterial pathogens infecting patients. The NIH recently experienced an outbreak of carbapenem-resistant K. pneumoniae that affected 18 patients, 11 of whom died. 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 [6].

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 at unprecedented depth. We now know that nine Mycobacterium species can cause tuberculosis, some of which, such as Mycobacterium bovis, require specific antibiotic treatments [6] Furthermore, the human microbiome project has shown that changes in the composition of microbial populations can drive disease [1].

1.2.4 NGS challenges and opportunities

Cost and speed: The cost and turn-around time for sequencing have been driven down by hardware advances. As a result, the cost for determining individual microbial genomes continue to fall (and now cost as little as $100 per sequence) [6] with multi-hour turn-around. Both will continue to improve. The justification for NGS will become increasingly apparent, starting with hospital patients who develop difficult-to-treat or life-threatening infections that already prove very costly to the system.

Informatics: NGS provides large datasets that require extensive bioinformatics for sequence analysis, from assembly to annotation. Furthermore, data presentation and distillation of clinical action are highly challenging. Because of the wide spectrum of microbiology labs in terms of technical sophistication, addressing the informatic challenges associated with NGS will be critical for widespread adoption of the technology.

Mechanism: Increasingly, NGS has been use to measure 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 and ?Ribo-Seq? for translation) or even molecular biophysics (for example, parallel analysis of RNA structure (PARS), and global mapping of DNA-DNA interactions using proximity ligation coupled with deep sequencing (Hi-C). Such methods could - in principle - also be applied to infectious disease.

1.3 Contributions and outline of this thesis

(1) NGS can be used to isolate infection-derived DNA fragments can be detected in human blood. (2) These measurements are clinically useful. (3) Applications can be used to manage this information. (4) NGS can also be used to understand 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 [1]. 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 [10]. 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 [3], molecular coating by NGS has been extended to organ transplants [2] (which can be thought of as genome transplants) and cancers [8]. The in the latter two cases, specific mutations can be used to resolve donor- or cancer-derived DNA fragments in circulation.

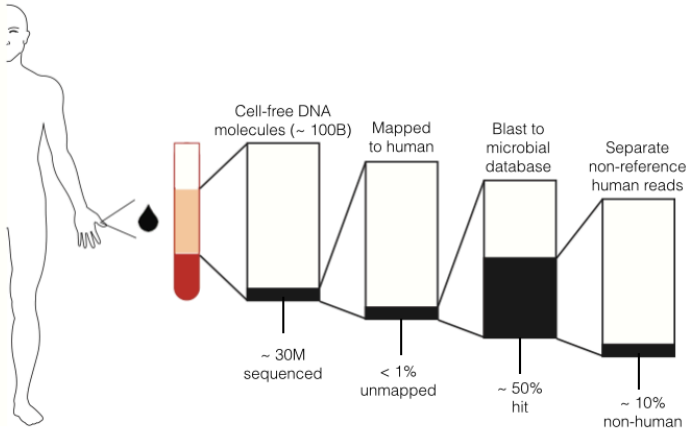

Figure 2.1: Isolation of non-human cell-free DNA.

2.2 Pipeline for capturing microbiome in cell-free DNA

Non-human (microbial, fungal, and viral) derived cell-free DNA fragments can purified from blood and counted using NGS suggesting that non-invasive detection of infectious disease from cell-free DNA may be possible. In part, this would be accomplished as described [2]: map short reads to the human genome, subtract all human-derived fragments, use BLAST to align the residual reads to a database of microbial genomes, and employ algorithms to reduce ambiguity in these alignments [13]. Two challenges frustration this vision. The biochemical challenge is primarily driven by the fact that non-human cell-free DNA is found at trace abundance in blood, comprising $< 1\%$ 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).

The second challenge is to organize and present the information at scale, across large clinical cohorts, such that it is intuitive for researchers and clinicians. 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 [7] 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

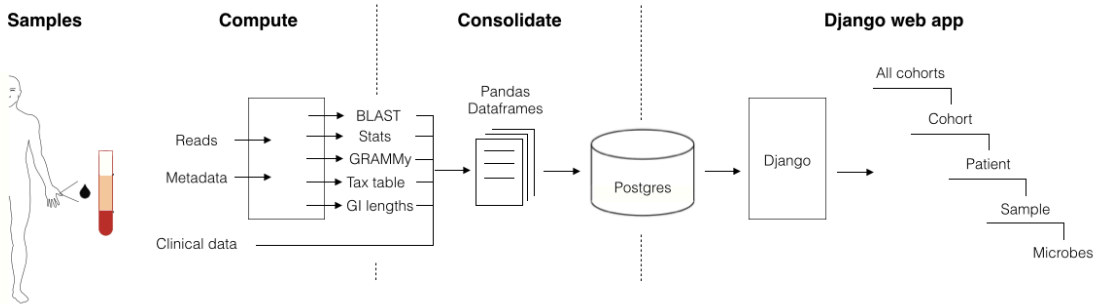

Figure 2.2: Django application for infectome data.

described previously, using several algorithms to compute relative abundance of microorganisms in each sample [2]. 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 [9]. 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

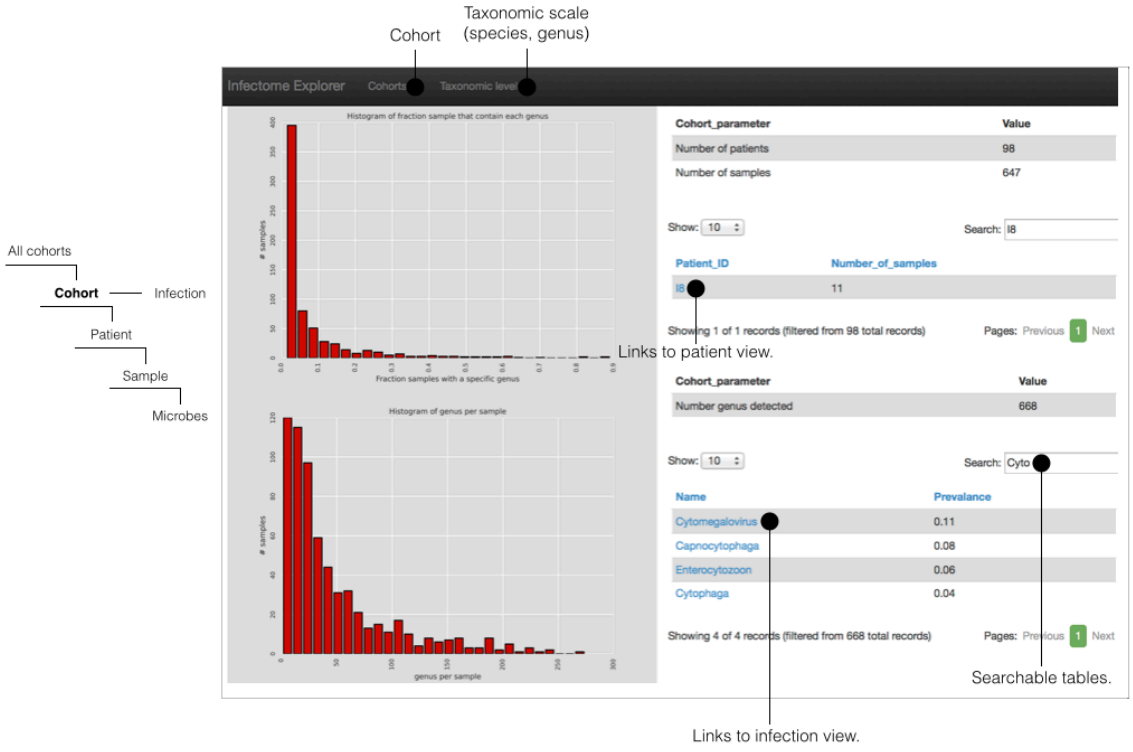

Figure 2.3: Cohort data

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

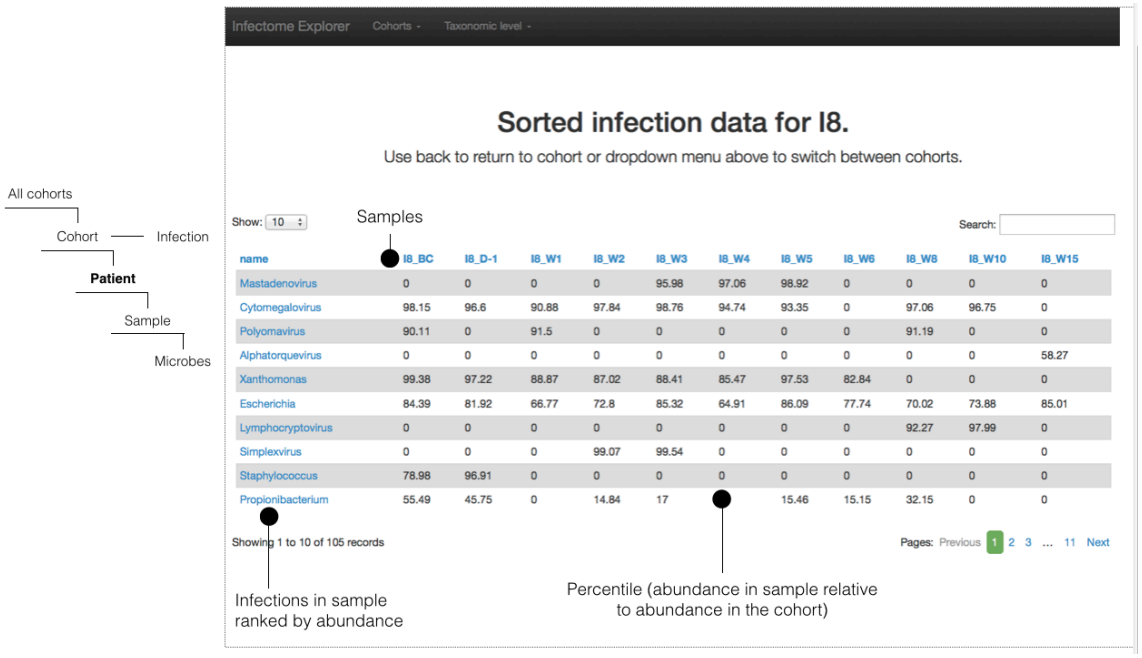

Figure 2.4: Patient data

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 2.4) 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

Figure 2.5: Infectione timeseries and coverage

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 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

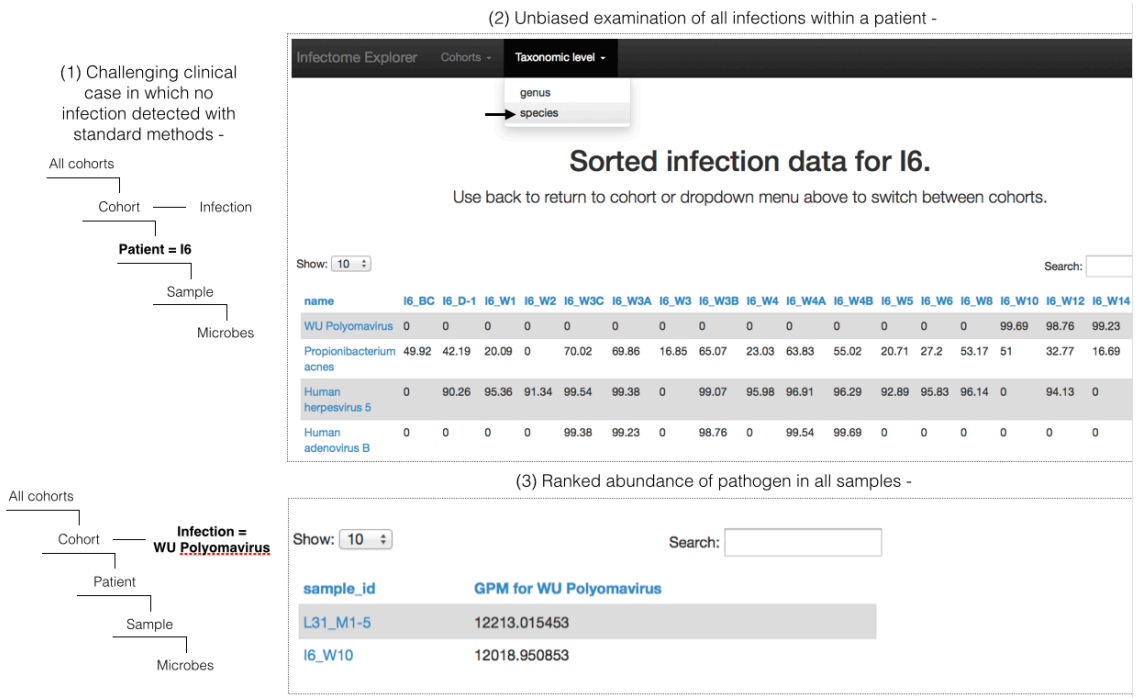

Figure 2.6: Clinical use of infectome application

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 [12]. 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 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

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 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

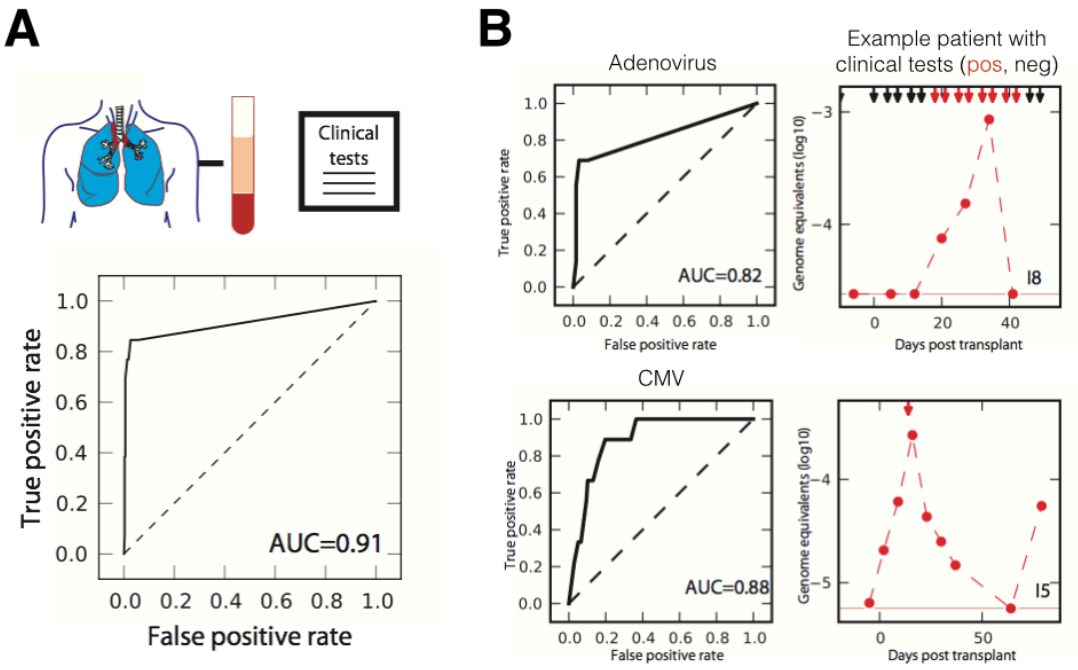

Figure 3.1: Clinical correlations on viruses

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 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

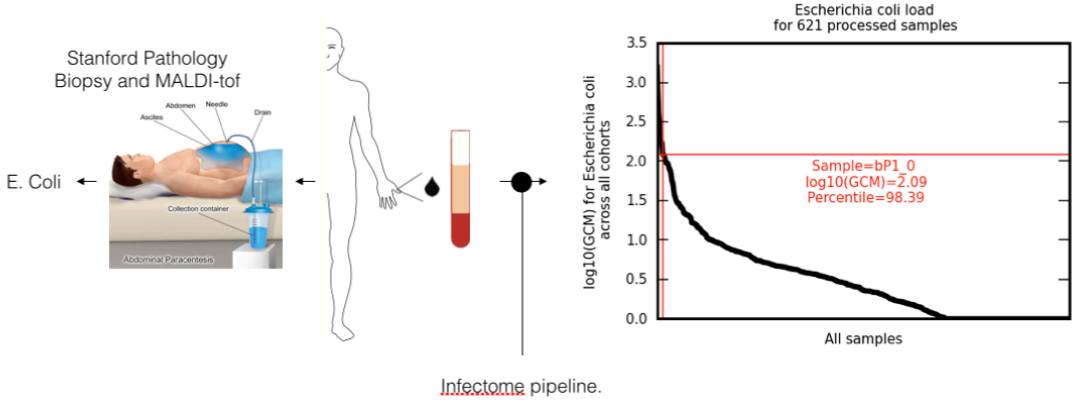

Figure 3.2: Clinical correlations with deep tissue sampling

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

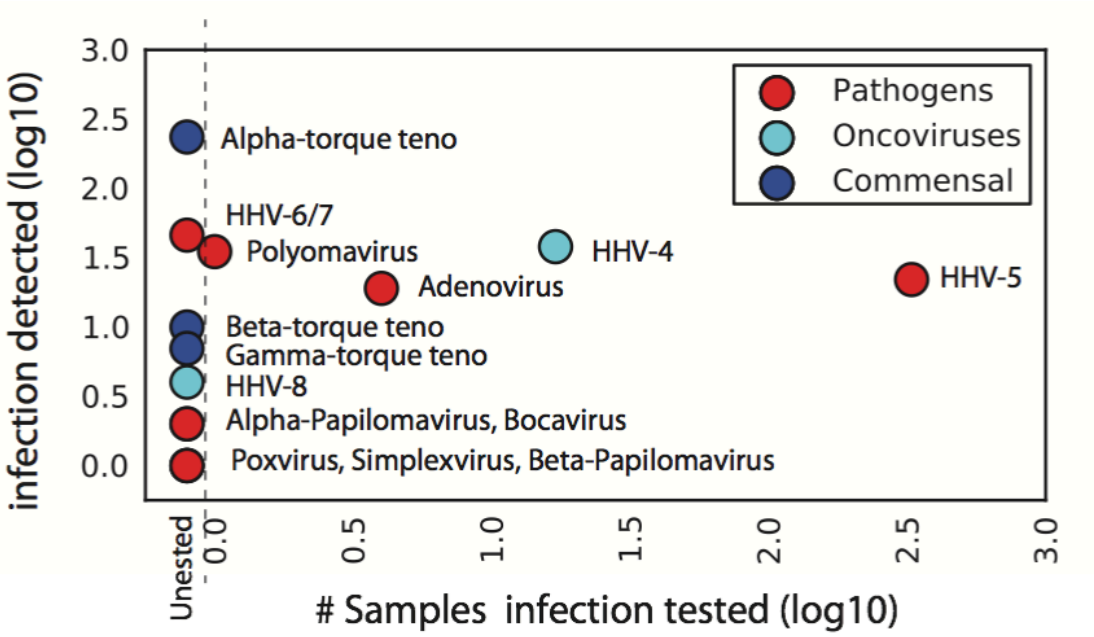

Figure 3.3: Clinical correlations on viruses

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.

Bibliography

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

[2] 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.

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.