COUNTING 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
February 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

Infectious diseases have a profound impact on humankind, influencing the course of wars and the fate of nations. The case for next-generation sequencing (NGS) in modern clinical microbiology is becoming increasingly clear. A flood of recent studies have shown how this powerful technology can address pressing modern problems in infectious diseases, including resistance, rare infections, and outbreaks.

In this thesis, we show that molecular counting of pathogen-derived cell-free DNA is a new diagnostic paradigm for infectious disease. We built a pipeline for counting pathogen-derived cell-free molecules in human plasma and a web application presenting the resulting data. We applied these tools to thousands of clinical samples collected from hundreds of patients at Stanford hospital. We further processed thousands of clinical test records in order to show that this method can be broadly applied for non-invasive monitoring of viral, bacterial, and fungal infections in deep tissues. Finally, we show that unbiased pathogen monitoring using this technique can track infections that escape hypothesis-centric clinical testing.

After demonstrating this new diagnostic application of NGS, we show how NGS technology can be used to understand infectious disease mechanism. We developed a pipeline for counting of sequencing reads derived from RNA-protein interactions *in vivo*. We show that this method (CLIP-seq) can be applied to viruses that have infected human cells and use it to reveal novel interactions between the HCV genome and human protein PCBP2. In a follow-up study, we applied the method to HERV-K, an endogenous retrovirus. We showed that human embryo development occurs in the presence of retroviral products, which protects the embroyo from exogenous infection while exerting regulatory function through interaction with human mRNAs.

We highlight three separate ways to validate results from mechanistic CLIP-seq experiments, including comparative analysis, replicate matching, and functional studies. We also developed a highly multiplexed RNA-protein interaction assay that is compatible with the scale of CLIP-seq experiments and far exceeds the throughput of common biochemical assays. We applied this technology to a model RNA-protein interaction (histone stem-loop and SLBP), recapitulating two decades of biochemistry in a single experiment while also revealing novel features of the interaction.

In summary, we built two computational tools that apply molecular counting to infectious disease diagnostics and mechanism. For validation of these results, we developed novel microfluidic tool for high-throughput biophysical measurments.

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 [11]. 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.

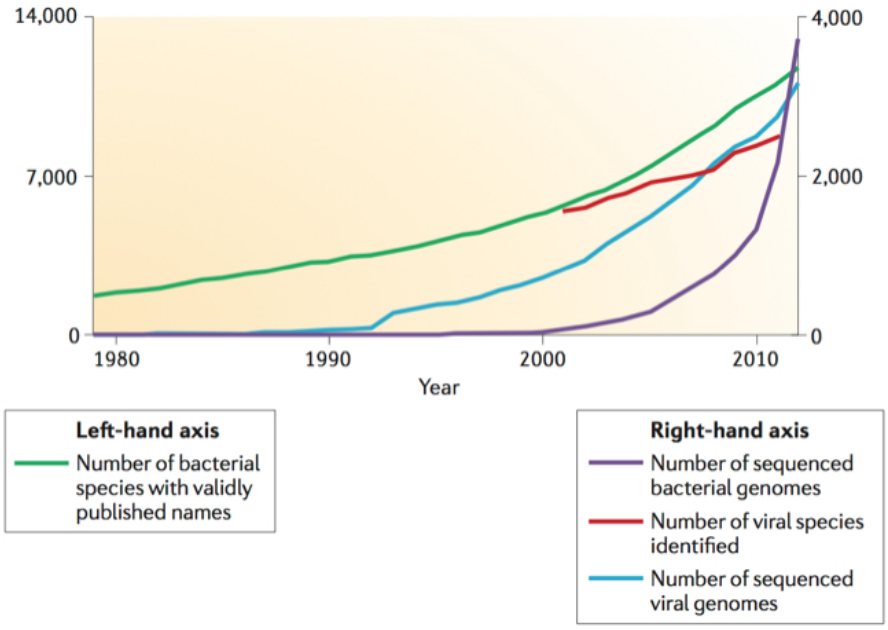

Figure 1.1: Rapid growth in the identification of micro-organisms.

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 [14]. 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 Kaposi's 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) [14].

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) [11] 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 [3].

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 [3]. 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 [15]. 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,

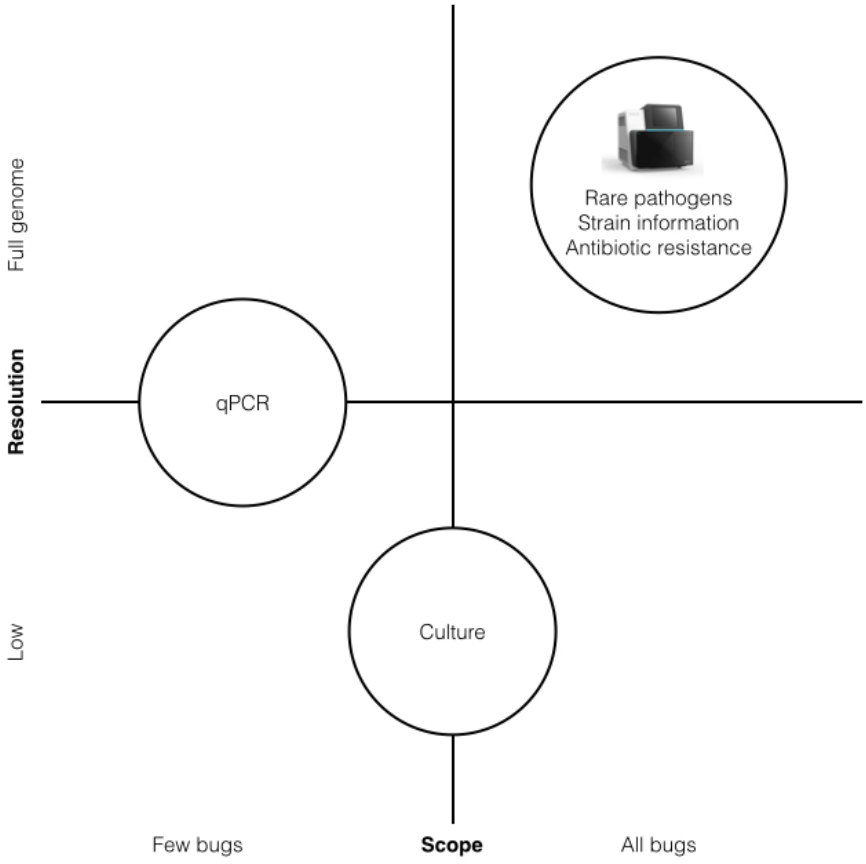

Figure 1.2: The trade-off between scope and resolution.

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 [3].

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 [34]. 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 [3].

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 [29]. (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 [42]. 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 [15]. 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 [3]. 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 [3].

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 [15]. 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 [3].

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 [15]. Furthermore, the human microbiome project has highlight the importance of assaying microbial populations [8]. 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 [15] 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) [34]. Many of these methods could also be applied to the study infectious disease.

1.3 Contributions and outline of this thesis

1.3.1 Molecular counting for diagnostics

In this thesis, we show that molecular counting of pathogen-derived cell-free DNA is a powerful diagnostic. We built a pipeline for counting pathogen-derived cell-free molecules in human plasma and a web application presenting the resulting data. We applied these tools to thousands of clinical samples collected from hundreds of patients at Stanford hospital. We further processed thousands of clinical test records in order to show that this method can be broadly applied for non-invasive monitoring of viral, bacterial, and fungal infections in deep tissues. Finally, we show that unbiased pathogen monitoring using this technique can track rare or un-expected infections that now escape hypothesis-centric clinical testing.

1.3.2 Molecular counting for mechanism

After demonstrating this new diagnostic application of NGS, we show how NGS technology can be used to understand infectious disease mechanism. We developed a pipeline for counting of sequencing reads derived from RNA-protein interactions *in vivo*. We show that this method (CLIP-seq) can be applied to viruses that have infected human cells and use it to reveal novel interactions between the HCV genome and human protein PCBP2. In a follow-up study, we applied the method to HERV-K, an endogenous retrovirus. We showed that human embryo development occurs in the presence of retroviral products, which protect the embroyo from exogenous infection while exerting regulatory function through interaction with human mRNAs.

We highlight three separate ways to validate results from mechanistic CLIP-seq experiments, including comparative analysis, replicate matching, and functional studies. We also developed a highly multiplexed RNA-protein interaction assay that is compatible with the scale of CLIP-seq experiments and far exceeds the throughput of common biochemical assays. We applied this technology to a model RNA-protein interaction (histone stem-loop and SLBP), recapitulating two decades of biochemistry in a single experiment while also revealing novel features of the interaction.

Chapter 2

Infectome pipeline

2.1 Cell-free DNA

In 1947, Mandel and Metais first observed that human blood contains circulating cell-free DNA molecules. These fragments enter blood as the detritus of dead cells and circulate with short (15 minute) half-life as nucleosome-protected fragments [33]. Methods of molecular counting (notably NGS) have taken advantage of this phenomenon, as the abundance of cell-free DNA species may correlate with human health. Their greatest value has been to measure the proportion of foreign genomes within an individual. Cancer-associated mutations can be used to determine the progress of disease [30], fetal DNA assayed by molecular counting can be used to detect aneuploidies (such as Down syndrome) [10], and donor-organ derived DNA fragments can be monitored as marker of rejection following transplantation [9]. One critical advantage in all cases is that these measurements are non-invasive.

The ability to resolve foreign genomes in blood presents an opportunity for infectious disease monitoring. The broad case for NGS in infectious disease diagnostics is already clear, particularly considering the growing importance of human microbiome on human health [8]. The application of NGS to micro-organism derived cell-free DNA fragments presents two new opportunities for infection monitoring: (1) Translocation of micro-organisms from body sites, such as the GI tract, into the systemic circulation is known to occur and often has detrimental consequences,

including immune activation or, in extreme cases, septic shock [4]. Direct measurement of composition in blood may be used to monitor active microbial translocation due to pathological damange of organ integrity. (2) Cell death and percolation of microbial products into blood may indicate the presence of infections compartmentalized within body sites, including deep tissues that are otherwise hard to access.

Recent work has shown that micro-organism derived cell-free DNA fragments do indeed exist and can be counted using NGS [9]. In turn, a critical question is whether these micro-organism derived cell-free DNA fragments are indicative of pathology within specific organ systems. In principle, these measurements can correlated with independent clinical tests of infetion establish their clinical relevance.

2.2 Pipeline for the cell-free microbiome

Prior to establishing the clinical relevance of micro-organism derived cell-free DNA, we built a computational pipeline and application for processing and analyzing the data, respectively. The general strategy for cell-free DNA isolation, sequencing, and read assignment have all been well-described [9]. First, human-derived reads are subtracted computationally using a short-read alignment algorithm (e.g, Bowtie). This step is followed by alignment (e.g., BLAST) to a reference database that contains sequences from candidate pathogens (e.g., NCBI). Algorithms to reduce ambiguity in these alignments [43] may also be employed. In spite of this, four challenges must be acknowledged in pipeline and application design.

Large data: Alignment algorithms must contend with large amounts of sequence data. For example, the Illumina Next-Seq can now output >100 gigabases (Gb) of reads per 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 and 42 Gb for all nucleotide sequences in the National Center for Biotechnology Information (NCBI) nucleotide (nt) collection (as of January 2013).

Signal: In general, only a small fraction of NGS reads in clinical meta-genomic data correspond to pathogens [29]. This is particularly true in the case of cell-free DNA, as non-human cell-free DNA is typically < 1% of reads based upon initial

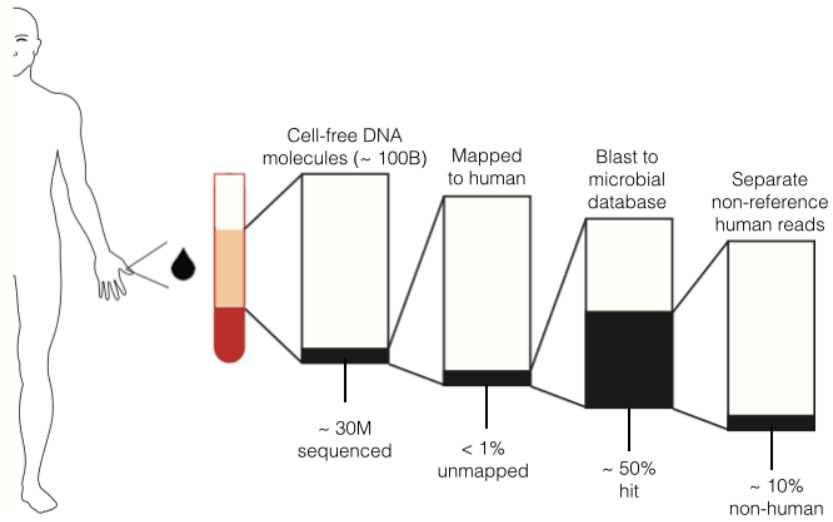

Figure 2.1: Isolation of non-human cell-free DNA.

human subtraction (e.g., mapping to the human genome). Furthermore, only a fraction of the un-mapped reads are derived from non-human sources, as most are either human-derived fragments that are not found in the human reference or do 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 like 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 classify potentially novel organisms [29]. Despite these technical advances, there will probably continue to be a gap between the availability of such data and the

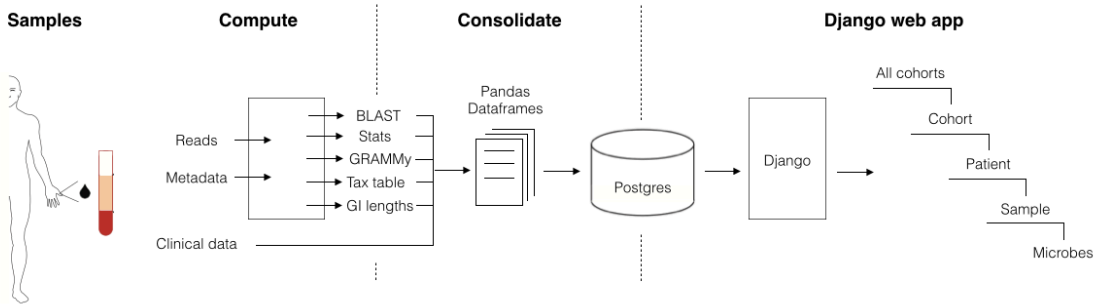

Figure 2.2: Django application for infectome data.

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 as well as data organization and visualization to aid with interpretation of the results. In order to achieve this, we built a pipeline comprised of cluster scripts that execute alignment, BLAST, and associated clean-up scripts [43]. We then build an application stack on top of this pipeline written in Python, using a Postgres relational database as the store for relevant pipeline-specific reference files (e.g., taxonomic table), outputs (e.g., BLAST results), and sample -meta data. We wrote the application using the Django web-development framework with the Matplotlib visualization library [18] and the Pandas library for data analysis, and Sqlalchemy for integration with Postgres (Figure 2.2).

We processed several thousands cell-free DNA samples for different clinical cohorts using this pipeline. Each cohort was 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 level resolution.

With this in mind, we designed our application for intuitive navigation of this multi-scale data and were influenced by a rich history of genomic data browsers, which have been useful tools for navigating genomic data since the early 2000s [31]. The Django application presents a series of web pages that reflect each level of organization in the data for intuitive browsing by researchers and clinicians.

The cohort page is the top level of organization in the application. It presents a

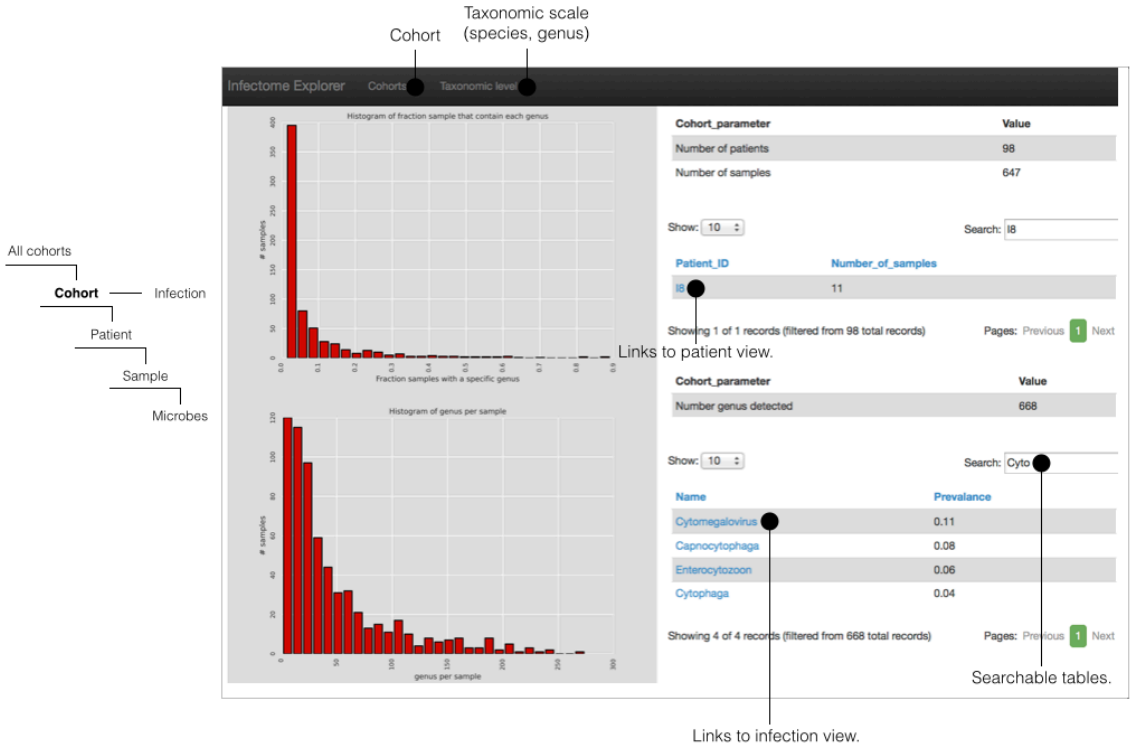

Figure 2.3: Cohort view in the infectome application.

table of patients, which is sorted by the number of samples per patient and provides a link to explore data for each patient in the cohort. It also provides cohort-level histograms that explain the incidence of each infection in the cohort (fraction of samples in which an infection is found) as well as the load per sample (the number of infections identified per sample). In addition, the cohort page provides a table of infections sorted 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). This make it possible to navigate the data in two ways: it is possible to take a patient-centric approach and examine specified patients (via the patient table) or an infection-centric approach (via infection table).

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

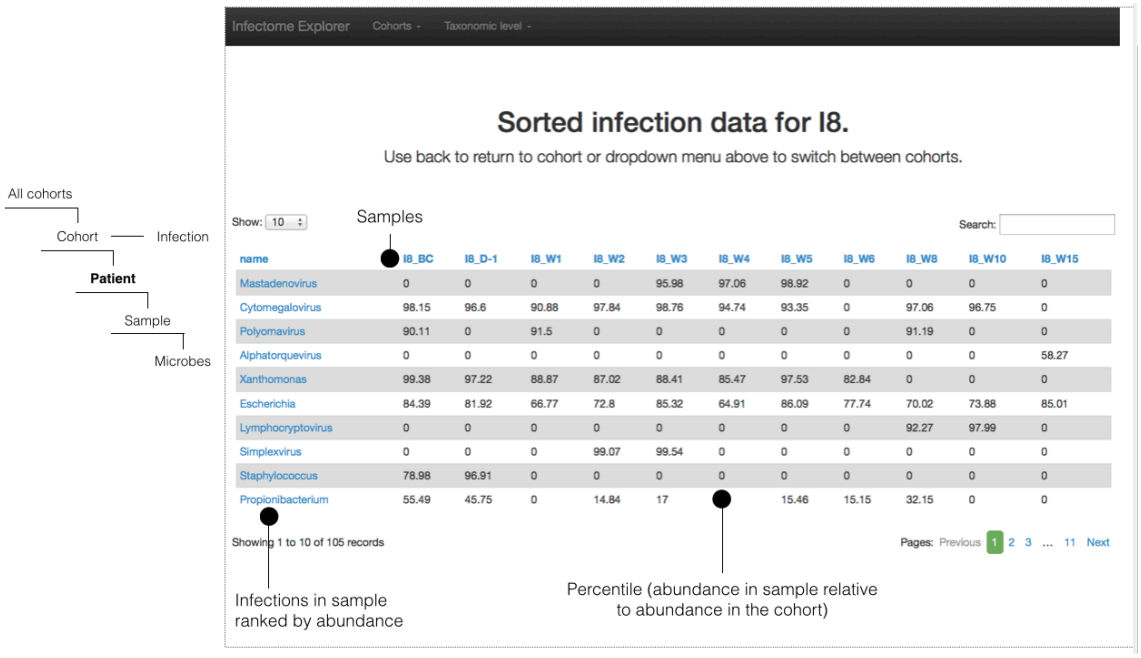

Figure 2.4: Patient view in the infectome application.

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 within each cell-free DNA sample.

From this measurment, we compute an estimate for absolute read counts per each identified genome. With this value, we then compute a coverage ratio between the infection and the human for that sample. We scale this value by 10^6, resulting in relative genome copies per million (gcm). In isolation, this value is reasonably intuitive: it indicates the number of genome copies for a given infection relative to sampled human-derived reads in that sample. A ratio of 1, for example, means that sampled organism genome copies is equivalent to sampled human genome copies.

For presentation, the raw gcm value may not be informative: a large value may

Figure 2.5: Infection view in the infectome application.

be use we convert this value into a percentile (Figure 2.4) with respect to the full cohort for that particular infection. The percentile value simply indicates the magnitude of each measurement relative to what was 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 time-series data for that infection as well as detailed information about read coverage across the microbial genome. Both measurements can provide greater confidence about 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.

Furthermore, coverage is computed directly from the raw BLAST data. The BLAST file provides an alignment of each read to a particular GIs (individual sequence record in the BLAST database). GIs are associated with NCBI taxIDs, which are unique identifiers for micro-organisms. We aggregate GIs present in the BLAST file by taxID. We concatenate GIs into a composite track for the associated taxID.

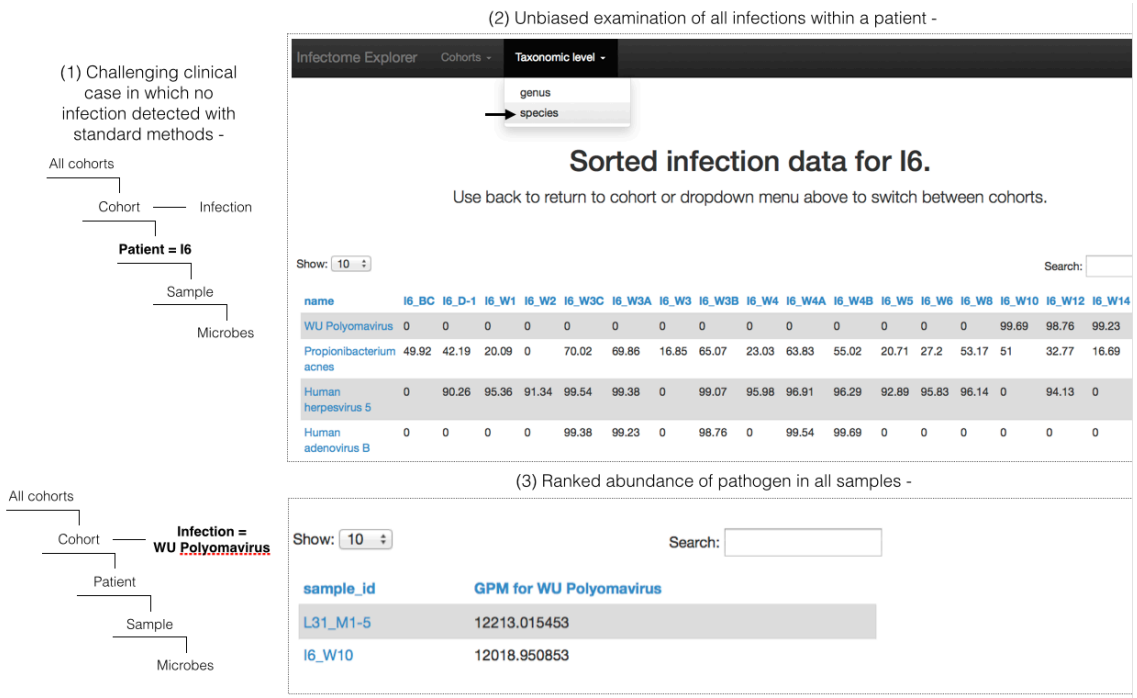

Figure 2.6: Clinical use of infectome application

We then compute the mapping position of all reads within the composite genome. Irregular coverage patters may be indicative of database contamination, as this may mean that all reads align only to a single GI associated with a given infection or may align to a narrow region within a given GI. Using data from a bone marrow transplant cohort patient (I8), we show both timeseries and coverage data (Figure 2.5).

To demonstrate clinical use for this application, we highlight the case of I6, a pediatric bone marrow patient with severe graft complications. We collected and processed longitudinal cell-free DNA samples over the course of post-transplant therapy. Using the patient-specific view, it was clear that I6 had a very high Polyomavirus load. Viewing the data at species-level, we can see WU Polyomavirus, a species that has been implicated in severe respiratory illnesses [21] (Figure 2.6).

Though the patient was tested for a different Polyomavirus (BK virus), those tests were negative. This situation is similar to a scenario recently described in the literature: NGS applied to cerebrospinal fluid within a deeply ill immunocomprised patient identified an exotic pathogenic bacteria, Leptospira, responsible for

encephalitis and informed successful treatment [42]. Unlike that case, I6 died prior to clinical intervention based upon this information. While it is not clear that WU Polyomavirus was responsible for the death of I6, it is clear that unbiased screening of potential pathogens in severe cases such as this one can reveal agents that escape clinical testing and serve as a powerful supplement to existing clinical assays.

Chapter 3

Clinical validation of the infectome

3.1 Organ transplantation

Organ transplantation is one of the pioneering applications clinical cell-free DNA based diagnostics. The ratio of recipient to graft-derived donor DNA, distinguished by SNPs that are specific to the recipient or 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, acute cellular organ rejection was marked by increases in the proportion of donor-derived DNA. In turn, this approach is less invasive and more accurate than traditional biopsies of the graft tissue [35].

Due to ongoing work on multiple transplants (heart, bone marrow, lung), the Quake lab had thousands of existing cell-free DNA samples sequenced. We processed these samples and isolated micro-organism derived cell-free DNA for each using the pipeline and application described previously (Chapter 2). Incidentally, these cohorts were well suited for evaluating the clinical utility of our measurements. Immunosuppressive therapies reduce the risk of graft rejection, but increase the susceptibility of recipients to infections. For example, infectious complications remain one of the most important causes of morbidity and mortality after lung transplantation, with cytomegalovirus infections (CMV) posing a significant threat.

As a result, frequent infection monitoring was performed on each transplant cohort. For the lung cohort alone, we collected over 35000 clinical measurements

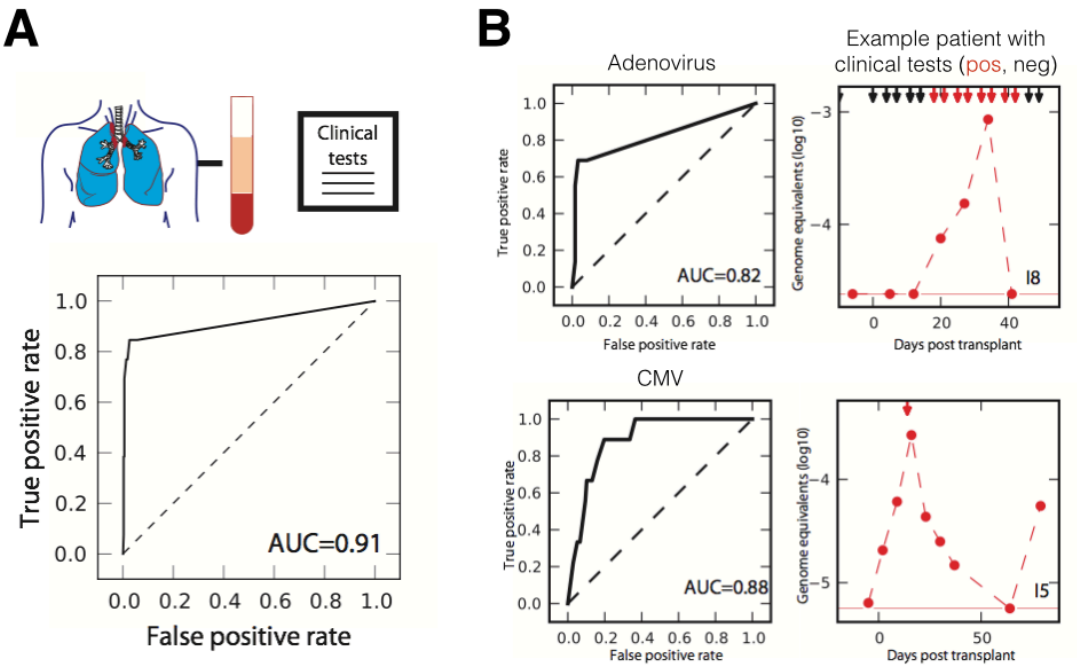

Figure 3.1: Clinical correlations on viruses.

of specific infections performed on 14 specimen types. The majority of these measurements were specific qPCR tests for CMV, a well-known risk factor in lung transplantation, and culture applied to bronchoalveolar lavage fluid, which is used to test for deep lung bacterial infections. We evaluated whether molecular counting of infection-derived genomes in cell-free DNA correlates with clinical test for infection.

Because there were a large number of clinical tests performed on CMV, it was a very good starting points for evaluating clinical utility of the cell-free measurements. We counted reads that map to the CMV genome for each lung transplant sample. We did this by first processing the BLAST data with an algorithm (GRAMMy) that computes relative abundance of each genome in sample based upon read mapping as well as genome size [43]. From this data, we compute a coverage ratio for each infection relative to human, which corrects for genome size as well as sampling depth. We found that this coverage ratio correlated well with clinical tests for infection, resulting in an AUC of 0.91 for CMV in the lung cohort (Figure 3.1).

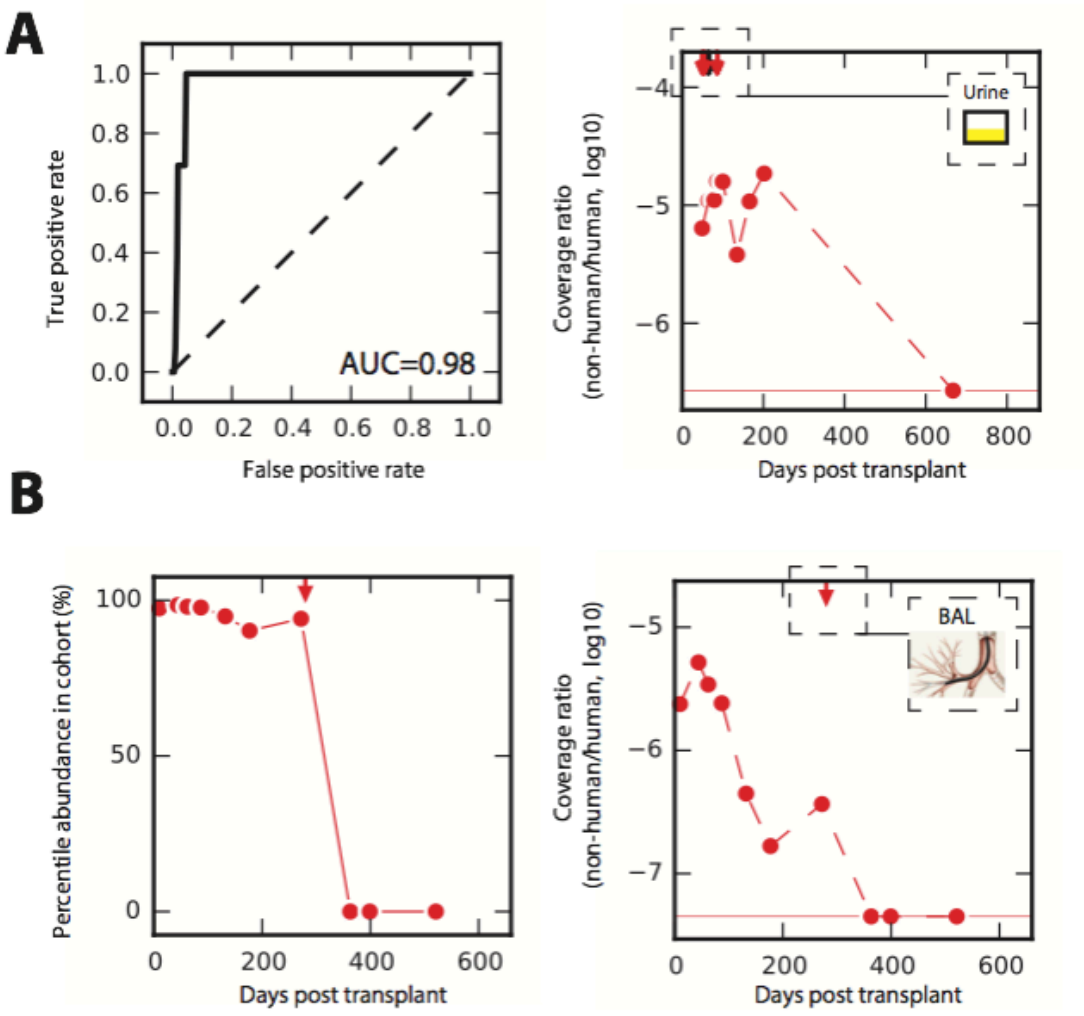

Figure 3.2: Clinical correlations on bacteria and fungi.

We then performed a similar analysis on tests collected for the pediatric bone marrow cohort. Like lung, clinical testing for CMV was common. Because of the elevated risk in the pediatric cases, this cohort also included regular screens for additional viruses, including Adenovirus (a community-acquired respiratory infection that can cause graft loss in transplant recipients and poses a particularly high risk for paediatric patients). We saw a correlation between infection-derived cfDNA and positive clinical tests in timeseries data, with elevation in signal observed when positive clinical test results were recorded (Figure 3.1). Cohort-level with ROC curves had AUC values of 0.82 and 0.88 for Adenovirus and CMV, respectively.

The performance on viruses is encouraging, because the cell-free DNA samples used in this study were enriched: the majority of the reads were human, leaving few reads that could have been sampled from infections present in the blood. The fact that we observed correlation in the face of this limitation is remarkable. A far better signal should be achieved when human-depletion strategies are employed.

In addition to viruses measured in serum, we also observed an agreement between cell-free DNA measurements and fungi and bacteria detected in other body fluids, including *Klebsiella Pneumonia* infections detected via urine culture (ROC = 0.98) and *Aspergillus Niger* infection detected in BAL (Figure 3.2). In the case of *Klebsiella*, there were 687 negative results and 13 positive results. All positive test results are in L24, and (except 1 for BAL) and are detected in urine, resulting in a consistent timeseries in our data that agrees with clinical results. Matched sample-tests data for *Aspergillus Niger*, which has 481 negative results and 2 positive results in L17 (1 BAL and 1 sputum) with the single test for BAL is shown. The percentile measurement highlights that measurements in L24 are large relative to the cohort, which is consistent with the positive clinical test results in this patient.

We also evaluated the performance on classes of pathogen as well as sample type. The performance of bacterial and fungal correlations depended both on the infection type as well as the body site queried. In general, we observed better performance for body sites that have tighter coupling to blood (Figure 3.3). We also found that the assay performance depends on the level of the normal background signal. For example, test performance for the most commonly cultured bacteria

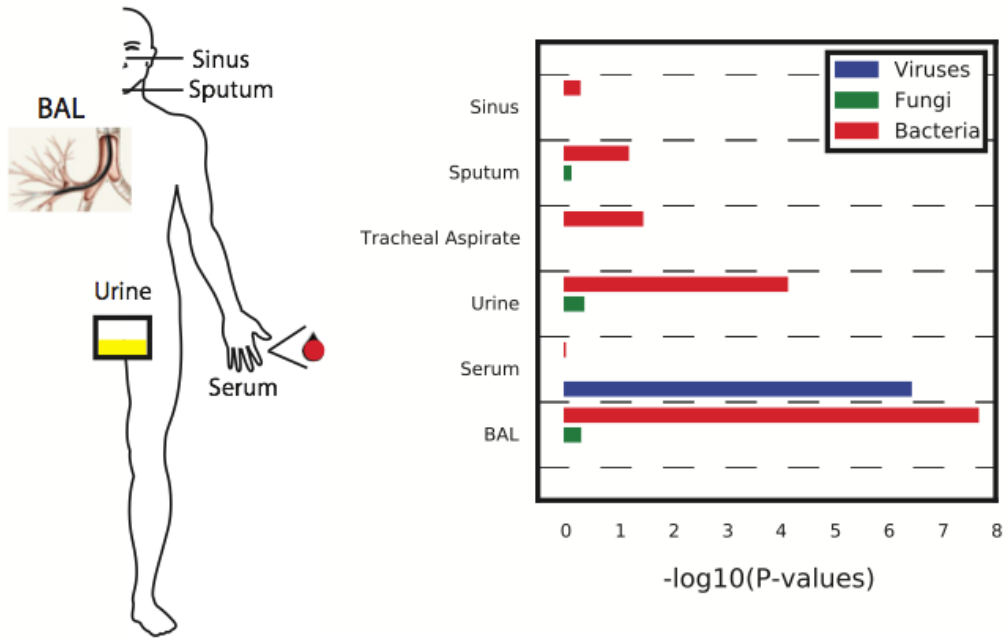

Figure 3.3: Clinical correlations across sample types.

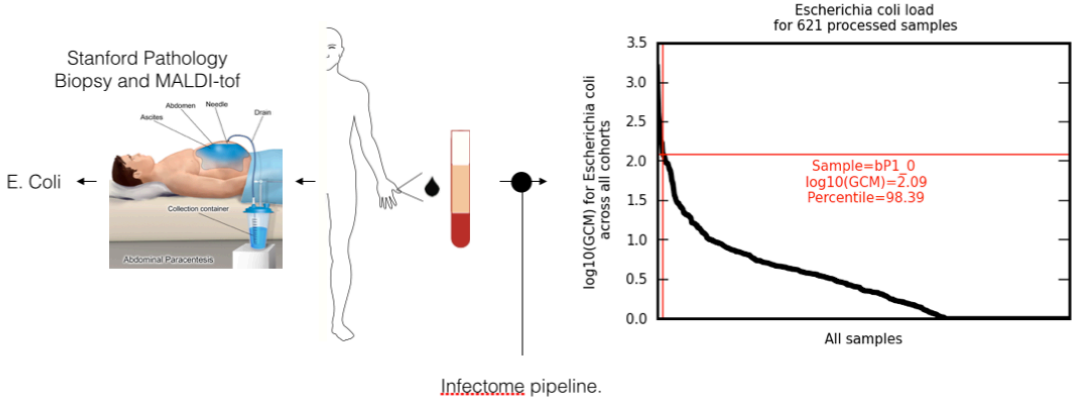

Figure 3.4: Clinical correlations with deep tissue sampling

(Pseudomonas), detected in over 80% of the patient samples was poor (AUC 0.62) in contrast to test performance for the most commonly detected viral pathogen (CMV), detected in only 6% of our patient samples (AUC 0.91). This highlights an important caveat in the ability of the plasma DNA sequencing for the detection of unusual abundance of commensal organisms (including Pseudomonas). Because they are part of the normal flora, many body site may contribute to the signal observed in blood and observe tissue-specific infection or dysbiosis due to pathology.

3.2 Deep tissues

We next examined whether molecular counting of infection-derived cell-free DNA correlates with tests for deep-tissue infections. This would be appealing, because invasive biopsies are risky, particularly in patients with compromised heath. We obtained blood samples from patients with deep tissue infections, which had biopsy performed and screened using MALDI-tof. We observed favorable correlation on the cohort in a pilot test of four samples. For example, on a patient with a gastrointestinal abcess that tested positive for *E. Coli*, we measured a very high (98%) percentile measurement of *E. Coli* in plasma for this patient relative to all other ($>$ 700) samples processed (Figure 3.4). In turn, this provides evidence that non-viral infections in deep tissues can be detected non-invasively via molecular counting.

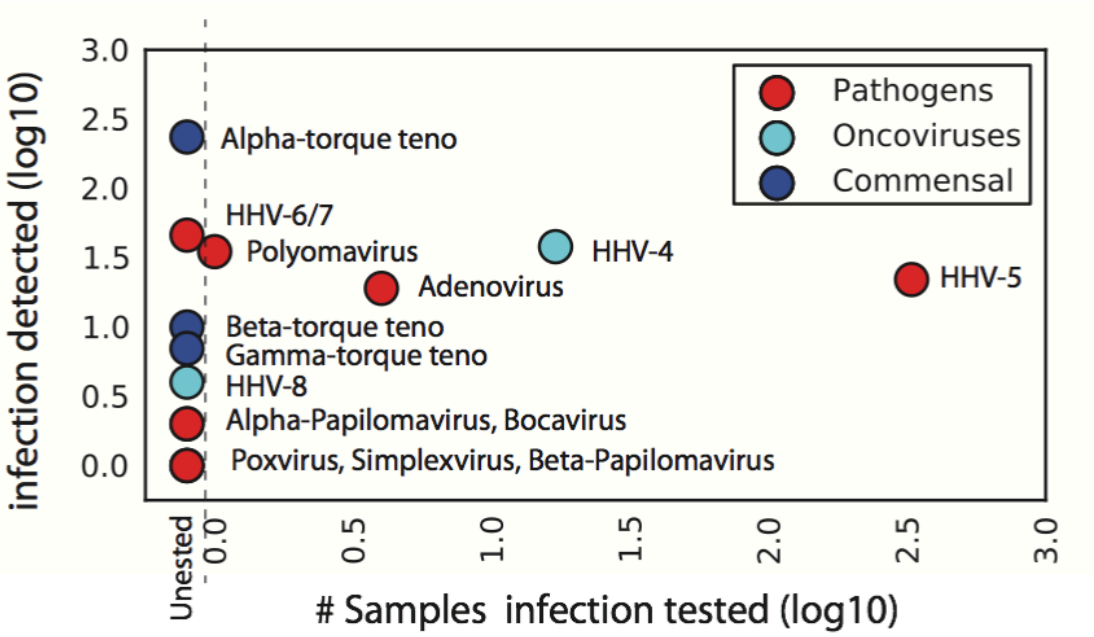

Figure 3.5: Clinical correlations on viruses

3.3 Untested infections

The benefit of unbiased molecular counting of infectious agents is particularly appealing, because it can indicate agents that currently fall outside the scope of clinical testing. We examined the merits of hypothesis-free screening by re-visiting our measurements relative to the recorded clinical data in the lung transplant cohort. In our data, we identified a host of viruses, ranging from well characterized pathogenic and onco-viruses to commensal torque teno viruses. The frequency of clinical testing for these viruses varied considerably, with frequent surveillance of CMV (Human Herpes Virus 5, HHV-5) relative to all other pathogens (Figure 3.5).

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 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

found that unbiased 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. With this mind, unbiased screening is a powerful compliment to existing, hypothesis-centric clinical tests for infection.

Chapter 4

The blood microbiome

4.1 Importance of the microbiome

The colon of an adult human contains approximately 10^{14} micro-organisms, which outnumber host cell numbers by up to two orders of magnitude and house at least 100-fold more bacterial genes than human [4]. The importance of the microbiome has become evident in part through studies of germ-free animals: these animals have reduced vascularity, digestive enzyme activity, and muscle wall thickness. Furthermore, numerous human studies have implicated changes in microbial composition (dysbiosis) with many diseases, such as obesity, celiac disease, type 2 diabetes, atopic eczema, asthma, inflammatory bowel disease, and chronic diarrhea [4].

4.2 Linking blood and body sites

The Human Microbiome Project defined the compositional range of the microbiome within healthy individuals [8]. Since this pioneering work, there has been great interest in studying changes in the microbiome across different physiological contexts. Pregnancy is one example, as preterm birth is a leading cause of neonatal mortality and can be driven by intrauterine infections. Until recently, it was thought that intrauterine infections originated in the lower genital tract and microbiota ascended

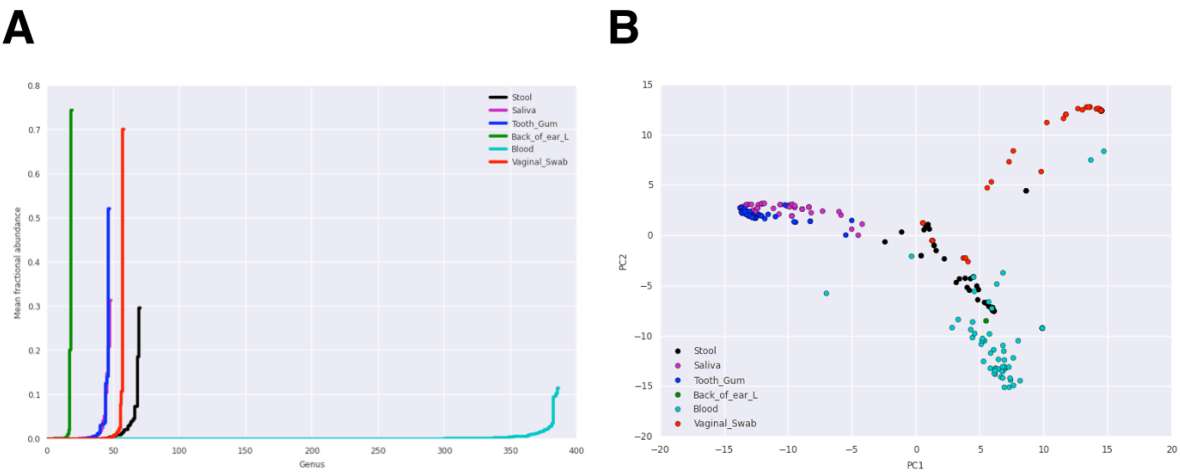

Figure 4.1: Composition of the blood microbiome

into the otherwise sterile womb [32]. Recent studies have shown that the microbiome changes during pregnancy according to hormonal and physical fluctuations [23]. In addition, the human placenta is not sterile, but rather harbors a unique microbiome with taxonomic composition similar to the oral cavity. The work suggested that transmission of micro-organisms from various body sites, potentially via blood, may help to seed (or alter the composition of) the placental microbiome [1].

Due to the apparent connection between the microbiome, blood, and fetal health, pregnancy is an interesting context for examining the linkage between the microbiome at body sites relative to blood. Recent work has shown that micro-organism derived cell-free DNA fragments can be purified from blood and counted using next-generation sequencing (NGS) [9]. Much of this material is likely to be the detritus of dead and apoptosed cells [33], which has leaked from tissues into the blood.

We examined the composition of these micro-organism derived cell-free DNA fragments recovered from blood relative to four commonly sampled body sites (Saliva, Vagina, Gut, and Gum) within a cohort of fourteen pregnant women enrolled in a clinical study at Stanford hospital. We performed shotgun sequencing of cell-free DNA collected from fifty eight plasma samples. For each sample, we performed temporally matched 16s sequencing on the four body sites.

We first performed descriptive analysis of the data by comparing the taxonomic

composition of blood relative to the sampled body sites at genus-level resolution. We discretized mean fractional abundance data for blood and all sampled body sites, resulting in a binary value for each genus. We then compared the genus detected in blood to genus detected in body sites. 58% of the genus detected in the body sites were also found in blood, while only 15% of the genus detected in blood were detected in any body site. This suggested that micro-organism derived material in blood originates from more than just the sampled sources.

We then evaluated the distribution of mean fractional abundances for each body site relative to blood. The 16s sampled body sites showed strong enrichment in particular genus that are known to be well-adapted to each niche [8]. 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. We transformed the data using PCA in order to determine the genus that most strongly explain the variation between samples for all tissues (Figure 4.1). The body sites cluster together in the transformed space, as expected, and blood occupies a distinct compositional niche relative to the sampled body sites. We examined genus that strongly contribute to the principal components in order to understand what genus distinguish blood from the body sites: as expected, genus that explain variation in blood relative to the sampled body sites (notably, Acidovorax and Cupriavidus) have the among the highest fraction abundance in blood, but are absent from the sampled body sites.

We next examined whether blood samples from each body site. Body-site specific micro-organisms provide a reasonable indicator for this. In turn, we computed specificity by discretizing the genus found in each body site and comparing discretized data for each body site to discretized data for all other sites (Figure 4.2). For this analysis, we used our sampled body site data as well as the metagenomic community profiles made available by the Human Microbiome Project [8], which contain 690 samples from 300 US subjects across 15 body sites.

In both cases, we computed a list of specific genus, which were only detected in single body sites for all samples collected in either study. We then asked whether these genus were detected in our blood measurements: we detected 57% and 45% of the site-specific genus in blood determined via HMP metagenomic community

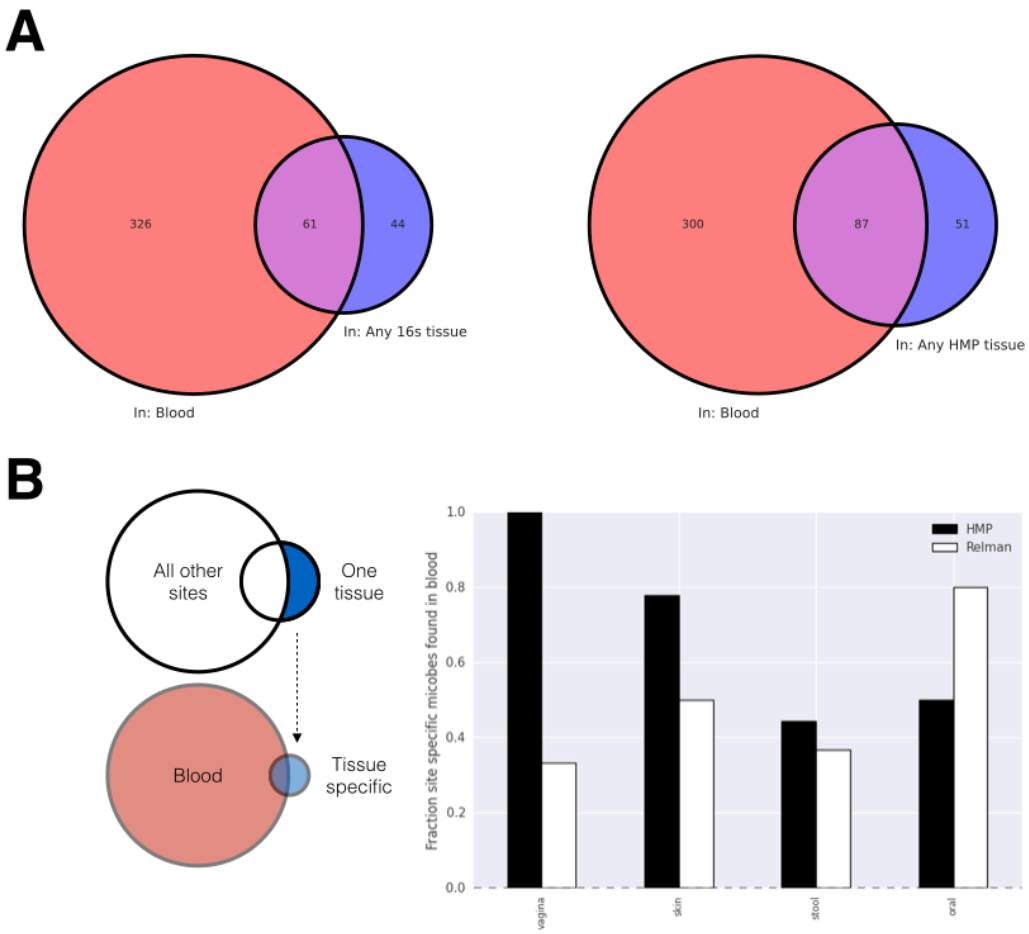

Figure 4.2: Detection of body site specific bacteria in blood.

profiles and the 16s data collected for this study, respectively. We found no significant difference between the abundance of site specific genus detected in blood versus theses un-detected, which argues against the possibility that technical limitations (e.g., under-sampling) explain the detection failure.

From the results of analysis, we obtained an assignment of body site specificity for each genus in our data; each genus was either assigned to a specific body site or not, the latter indicating that it was detected in more than one body site. We partitioned the genus detected in blood using this assignment and aggregated fractional abundance within each partition. This analysis indicated that blood is composed

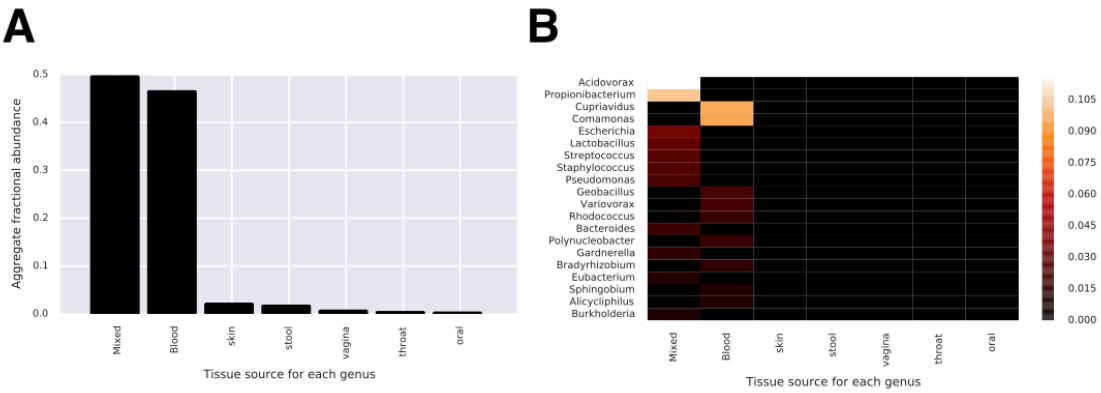

Figure 4.3: Likely sources for most abundant bacteria detected in blood.

primarily of micro-organisms with two likely source: genus derived from mixed sources, which cannot be traced to any specific tissue, and genus that are only detected in blood (Figure 4.3a). The 20 most abundant genus detect in blood (Figure 4.3b) were either exclusive to blood or originate from mixed sources.

4.3 Coupling between blood and body sites

We then asked if a physical relationship between genus found in blood and body sites can be established. Specifically, we examine whether the fractional abundance of genus at any body site is linked to its detection in blood. In order to do this, 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 compared the distribution of abundances at each body site based upon whether the genus was found in blood, expecting a difference in distribution if a body site was detectably coupled to blood (e.g., an elevation in abundance at a tissue when the genus was found in blood). However, found no signifiant difference between the body site fractional abundance with respect to detection in blood.

We further examined whether the composition of genus detected in blood can be modeled as a function of the sampled sites. We took a simple approaching, choosing a linear model. This assumes that each blood sample is a linear combination of the

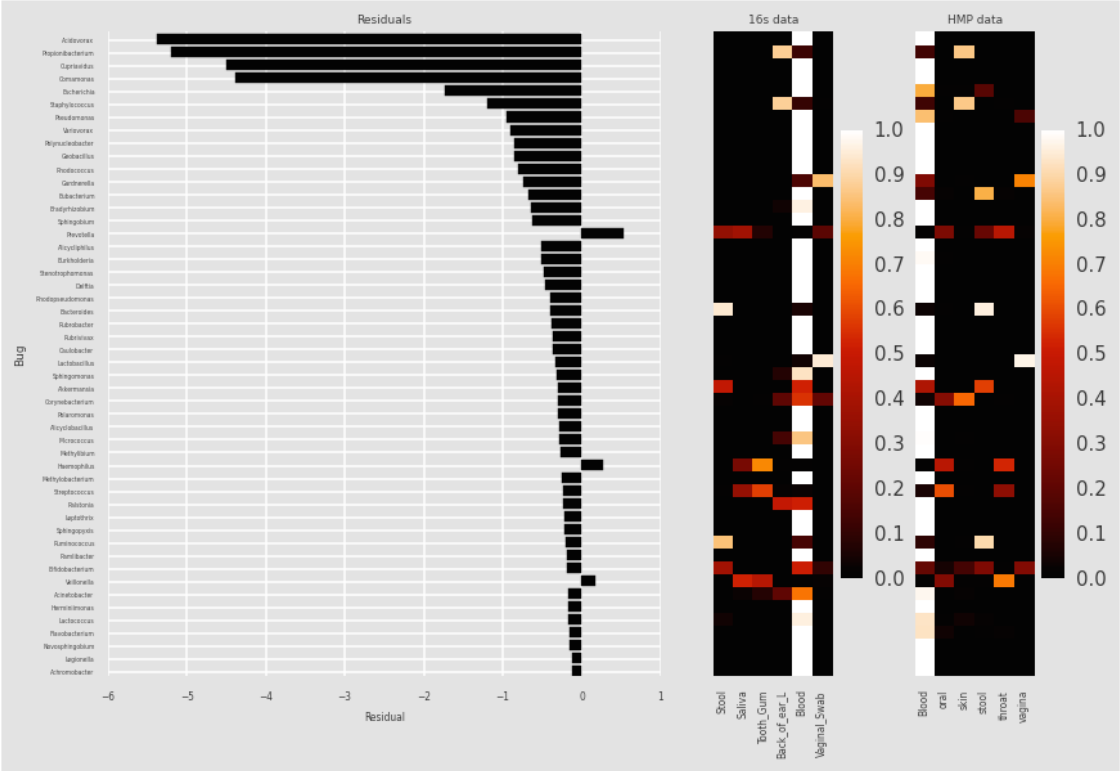

Figure 4.4: Residuals from linear regression.

genus at temporally matched body sites. We used a quadratic programming package in order to determine the mixing coefficients that minimize the squared error between the model output and the actual blood measurement subject to intuitive constraints (e.g., the coefficients must be greater than zero). After confirming the solver works correctly by recapitulating the correct mixing coefficients on simulated data, we applied this approach to all blood samples with matched body site data.

We found that the model performed poorly, using coefficient of determination (r^2) between the model guess and actual blood data as a measure of fit. However, examination of the the residuals was informative: intuitively, we found that the model fails because many highly abundant genus detected 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 samples from additional tissues or serves as an environment that amplifies specific genus non-linearly.

4.4 Summary

The abundance distribution of genus measured in blood differs from the sampled body sites. Body sites are dominated by few genus with high fractional abundance and have short tail of low-abudnance genus. This likely reflects the fact that specific genus are well-adapted to the niche provided by each body site [8]. In contrast, blood contains many more genus, but with long tail detected at trace abundance. 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. Yet, we also found that most abundant genus in blood were absent from the sampled body sites. This suggests that that blood is either dominated by genus contributed by body sites that were un-sampled or that blood may is a niche for these genus.

Analysis of site-specific genus provides evidence that each sampled body site contacts blood. However, we did not observe a significant relationship between the detection of a particular genus in blood and its abundance in any sampled body site. This may have several explanations. First, our sampling of micro-organism derived cell-free DNA fragments is limited by the high background of human derived signal. In turn, our sequencing depth of micro-orgaism derived fragments in blood may be insuficcient. Second, body site 16s measurements are normalized, resulting in a fractional abundance value for each detected genus. Yet, the microbial population sizes may differ considerably between body sites, meaning that fractional abundance may be poor indicator of the absolute contribution that a genus can make to blood. Finally, many body sites may be contributing similar genus to blood, obscuring the apparent linkage between abundance at any one site and the abundance in blood. This point is supported by our results from linear regression analysis: we detected many high abundance genus in blood that were not measured in any of the sampled sites. Thus, it is likely that additional sites contribute to blood.

We have shown that blood contains several-fold more genus than body samples sampled using either 16s or metagenomic sequencing. Each sampled tissue contributed to blood, though the most abundant genus detected in blood appear to originate from un-sampled tissues. The depth of sequencing employed in prior

studies has been sufficient to show a correlation between the abundance of micro-organism derived cell-free DNA and significant changes to the host, including deep tissue infection and immunocompetence [9]. Yet, this study suggests that depth of sequencing employed may be insufficient to establish a correlation between cell-free DNA and commensal micro-biome composition at the sampled sites. With deeper sequencing of blood along with more extensive profiling of the human microbiome, it may be possible to (1) explain the currently un-defined tissues of origin for many of micro-organisms detected in blood and (2) understand physical linkage between microbial abundance at body sites and blood.

Chapter 5

Sequencing to explore mechanism

5.1 Therapeutics

5.1.1 Antibiotics

While NGS has great promise for infectious disease diagnostics, therapies are required to eradicate infections once identified. In the late 19th century, pathogen-immune serum was used as a successful treatment against infectious agents. This approach encouraged scientists to develop chemicals that kill the specific pathogens, starting with Ehrlichś "magic bullet" against syphilis (arsphenamine) in 1910. Within two decades, a generation of scientists were working on antibiotics. As a result of these efforts, sulfa drugs were developed in 1936 and penicillin in 1943. Nearly all antibiotics in use today are compounds that were discovered during the 1940s to 1960s - the golden era of antibiotic discovery - or their derivatives. Most of these compounds were discovered by screening soil-derived actinomycetes, but natural product discovery became impractical due to the increasing difficulty of identifying new classes of antibiotics against the background of known compounds [25].

5.1.2 Antivirals

When antiviral drugs were first developed in the 1960s, they did not seem to be particularly promising, with a few exceptions. In response to the HIV/AIDS pandemic, however, the development of antiretroviral drugs markedly expanded the arsenal of available antiviral agents. By the mid 2000s, 37 chemicals (plus IFN- α in both pegylated and unpegylated forms) were formally approved for the treatment of viral infection [7]. At least half of these were intended to treat HIV infections and there are a similar number of compounds are under preclinical or clinical development, at least half of which were expected to reach the antiviral drug market.

Critically, these drugs largely target molecules required for viral replication. Antiviral strategies generally target viral DNA polymerase for the treatment of DNA virus infections, helicase/NTPase for the treatment of HSV, HCV, or SARS-CoV infections, IMP dehydrogenase for the treatment of HCV and some negative-strand RNA virus (for example, arena- and bunyavirus) infections, SAH hydrolase for the treatment of other negative-strand RNA virus infections such as Ebola and Marburg virus or RNA virus infections such as rotavirus, and RNA-dependent RNA polymerase for the treatment of other positive-strand RNA virus (e.g., flavivirus) infections [7].

5.1.3 The challenge

A central challenge in the fight against infectious disease is evolution: pathogens mutate in response to selective pressure imposed by treatment, resulting in an escalating arms-race between man and infection. Consider that RNA viruses exhibit extremely high mutation rates, orders of magnitude greater than those of most DNA-based life forms. Studies carried out to date suggest that many RNA viruses generate 10^{-4} to 10^{-6} errors per nucleotide, which is equivalent to approximately one mutation per genome per replication cycle [24]. Given the large population sizes observed in both experimental and natural infections with these viruses, every possible point mutation and many double-mutation combinations could theoretically be generated during each replication cycle within a population.

5.2 Mechanism

5.2.1 The case for mechanistic studies

Considering that pathogens evolve, the effectiveness of therapies today cannot be assured tomorrow. Indeed, there are now great concerns about the emergence of antibiotic resistant bacteria as well as viral strains (e.g., of HIV) that no longer respond to antivirals. In turn, there are been great emphasis on understanding the molecular principles that underlie infectious disease [15]. These molecular principles can be used to devise new therapeutic strategies.

5.2.2 The importance of molecular interactions

The study molecular interactions, particularly in the case of viruses, has been a productive way to devise new therapeutic strategies. This is particularly evident in the case of Hepatitis C Virus (HCV), a global health concern with 2 - 3% of the world?s population infected [41]. HCV is a positive-sense single-stranded RNA virus of the family Flaviviridae. The 9.6 kb genome contains a single open reading frame that is subsequently cleaved into 10 viral proteins and is flanked by UTRs.

The standard of care against HCV is a combination IFN / ribavirin, although many patients do not benefit from this treatment. With this in mind, is widely expected that in future small molecule drugs that target specific viral proteins that play essential roles in the viral life cycle (a.k.a. direct-acting antivirals) will replace IFN-based therapies. The approvals of two protease inhibitors (2011) and polymerase inhibitors (2013) were significant milestones in this regard.

Parallel efforts to discover novel viral targets have also been effective. For example, HCV interacts with numerous micro-RNAs (miRNAs), which are molecules predicted to regulate at least 60% of all human genes. One miRNA, miR-122, promotes HCV accumulation through direct interactions with the viral genomic RNA. Mechanistci studies have shown that miR-122 stabilizes the viral RNA by protecting the 5ŕterminus from degradation by the host exonuclease, Xrn-1 [41].

With this molecular knowledge in mind, agents that target miR-122 have been

used to treat HCV infection. In a recent Phase II clinical trial, miravirsen, an antisense locked nucleic acid molecule that binds to and sequesters miR-122, reduced serum HCV titers in treatment HCV-infected patients. At the highest doses used, HCV RNA became undetectable, but rebounded following completion of the 4-week course of miravirsen mono-therapy. A 12-week course of treatment is currently being tested to determine if patients can achieve sustained viral clearance [41].

Targeting host molecules, such as miR-122, may have numerous potential advantages, including a higher barrier to resistance, pan-genotypic activity and a wide range of druggable targets (whereas viral targets are limiting) [41]. Thus, a better understanding of the host-viral interaction networks that underlie translation and replication are likely to reveal novel targets for therapeutic intervention.

5.3 RNA-protein interactions

Considering the importance of molecular interaction networks for the development of new therapeutic strategies, we ask whether NGS-based strategies may be used to better understand infectious disease mechanism. Because HCV is a very well-studied model system, we first considered how NGS may be used to better understand RNA-viruses. Two points were clear: (1) RNA-protein interactions are central for the translation and replication of RNA viruses, with host proteins (e.g., ribosomes) critical for the production of viral products. (2) New NGS-based biochemical methods, such as CLIP-seq, make it possible to perform genome-wide RNA-protein measurements [22]. With these points in mind, we decided to build a computational pipeline for processing CLIP-seq data that can be easily applied to viruses.

5.4 CLIP pipeline

5.4.1 Philosophy

Considering the diverse reach of RNA-binding proteins (RBPs) in cell biology, substantial effort has been focused on methods for genome-wide interrogation of RNA

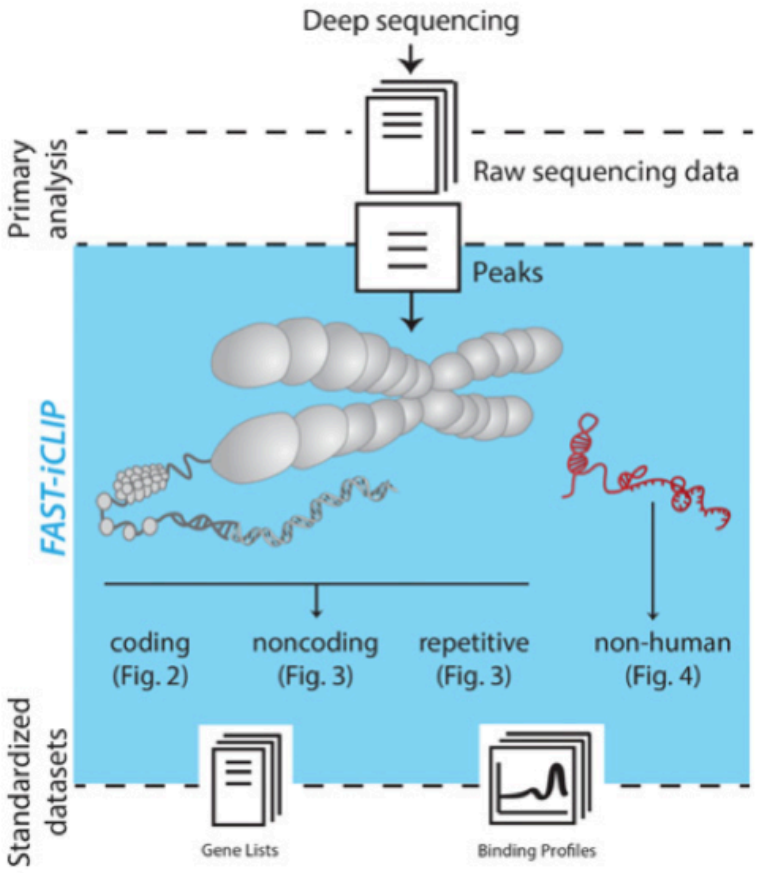

Figure 5.1: CLIP analysis workflow.

protein interactions using NGS. By stabilizing direct interactions in vivo combined with stringent purification steps, UV cross-linking immunoprecipitation and sequencing (CLIP-seq) enables specific isolation of an RBPś RNA-binding sites for NGS [22].

Much of the pioneering work has been focused on well-studied proteins and the protein-coding transcriptome, leading to numerous important advances on both methodological (PAR-CLIP, iCLIP, and BrdU-CLIP) and computational fronts [12]. These results have spurred broader interest in CLIP, particularly with respect to the interactomes of noncoding RNAs or diversity of viruses and microbes that impinge on human health [8]. While RNA protein complexes such as the ribosome and

spliceosome are well-studied, a vast and enigmatic repertoire of noncoding and nonhuman RNA-protein interactomes await further characterization.

Yet, extending CLIP across many RBPs is challenging for at least two reasons: (1) The sample preparation protocol is inefficient and time consuming and (2) informatic methods are not easily implemented or generally applicable to any RBP, particularly if RBP targets are not obvious *a priori*. To put these challenges in context, the CLIP workflow can be thought of as a stack of tasks - starting with NGS biochemistry, followed by informatic transformations of the resulting data, and finally protein-specific questions or analyses. Because the specificity of work increases as one moves across the stack, we sought to address common challenges to any CLIP investigation by improving the efficiency of sample preparation, extending the intermediate analysis to include a diverse set of user-definable transcriptomes (protein coding, non-coding, non-human, etc), and also standardizing data format output such that comparisons between RBPs are straightforward (Figure 5.1).

5.5 CLIP pipeline applications

Prior to testing our pipeline on non-human genomes, we first tested it on human targets. We focused on RNA helicases, which are conserved enzymes that use the energy of ATP to remodel RNA secondary structures and RNA-protein complexes.

5.5.1 Application to DDX21

The nucleolar helicase DDX21 is required for ribosome biogenesis and pre-rRNA processing, but the specific mechanism underlying this critical role remains unknown. To explore this, ChIP-seq in HEK293 cells was first performed, revealing DDX21-binding at promoters for genes involved in the ribosomal pathway. DDX21 knockdown decreased the steady-state levels of transcripts originating from DDX21-bound promoters, indicating that DDX21 associates with and positively regulates transcription of Pol I- and Pol II-dependent ribosomal genes [6].

The next question was whether DDX21 associates with RNAs directly involved

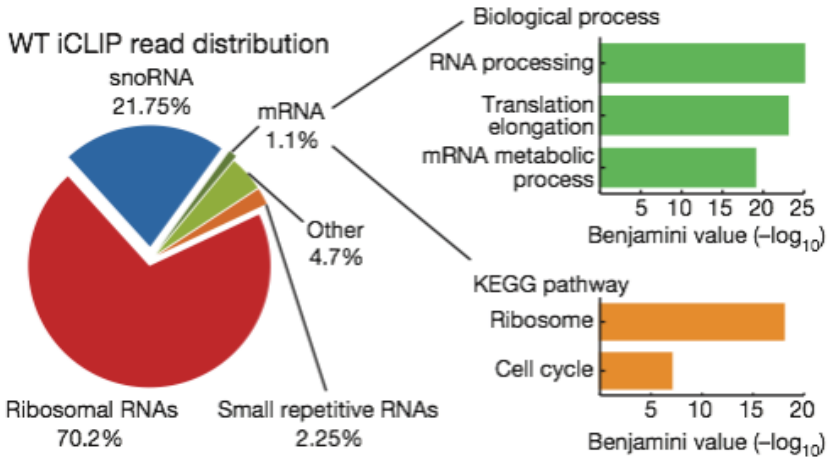

Figure 5.2: Bound classes of RNAs to DDX21.

in pre-rRNA processing. There were two clear candidates: (1) rRNA may be bound by DDX21. rRNA must be both cleaved and processed with chemical medications, such as pseudouridylation. These chemical modifications aid in the formation of ribosome complexes and may aid translational efficiency. (2) Chemical modifications to rRNA are made - in part - by snoRNAs, a specific class of small non-coding RNAs that guide protein complexes (e.g., enzymes) to specific modification sites on rRNA.

Our CLIP pipeline was well-suited to this question, because it was designed to analyze non-coding RNAs, including both rRNA and snoRNAs. In turn, we performed tandem purification iCLIP and processed the data, which partitions the data by non-coding RNA category. We found that DDX21 interacts with a diverse set of RNAs, of which rRNA and snoRNAs were most highly represented (Figure 5.2).

For the mRNAs bound, Gene ontology term and KEGG pathway analysis linked these mRNAs to ribosome function. This provides some evidence that the bound targets are bone-fide, rather than noise, as their is functional consistency between these genes and the predicted target pathway of DDX21. However, this is not enough. There are at least two ways additional ways to validate the results: (1) In order to convince ourselves that the CLIP signal is not spurious or noise, we

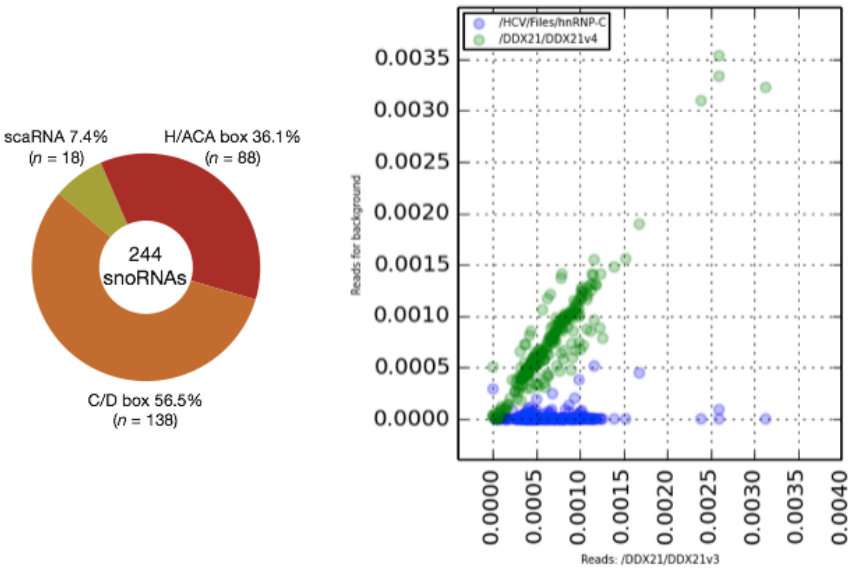

Figure 5.3: The snoRNA binding profile of DDX21.

can ask whether the observed binding profile is unique to DDX21 relative to other CLIPed proteins. (2) Most importantly, we can perform functional studies based upon, and to validate, the hypotheses generated by the CLIP data.

Because the CLIP pipeline generates a consistent data format for each protein processed, it is relatively easy to compare results for any class of non-coding RNA. We would like to compare the number of reads mapping to a particular transcript between experiments. Of course, experiments have different degrees of sequencing depth. To correct for this, we can divide the number of hits to a particular gene by the total number of mapped reads, resulting in a normalized count ratio that is more reasonable to compare. We perform this comparison for each bound snoRNA gene between the DDX21 CLIP dataset and hnRNPC, another RNA-binding protein on which iCLIP was performed [45] (Figure 5.3).

A scatter plots of normalized counts comparing DDX21 replicates along with DDX21 and hnRNPC indicates two key points: First, snoRNAs are similarly bound in DDX21 replicates. Second, these same snoRNAs are not also bound by hnRNPC. In turn, the snoRNA binding patters appears to be both reproducible as well as specific

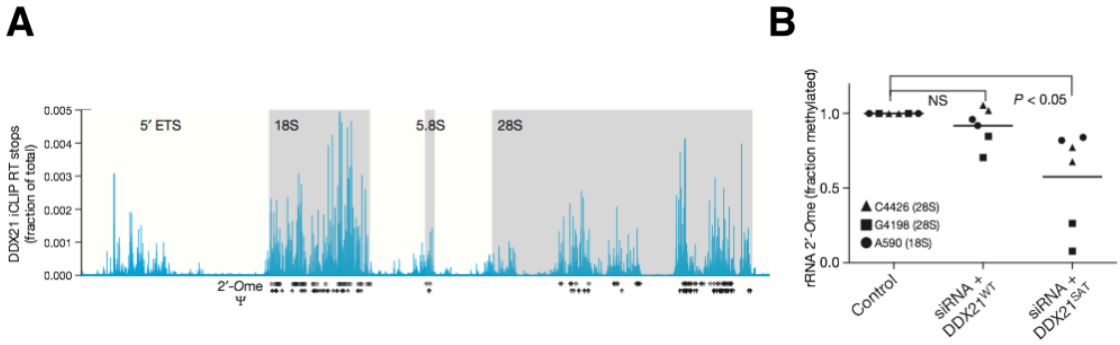

Figure 5.4: The rRNA binding profile of DDX21.

to DDX21. Yet, this is not sufficient to make strong claims about function. Using immunoprecipitation, we found that DDX21 cross-links to NOP58, fibrillarin and dyskerin, which are all protein components of snoRNP complexes. These data place DDX21 within snoRNAP complexes. These complexes perform enzymatic modification of rRNA, which suggests a testable hypothesis: if DDX21 is a key element of the snoRNP, then DDX21 knock-down should inhibit rRNA modification.

This can be tested by assaying specific rRNA modification expected to be controlled by DDX21-dependent snoRNP function. In turn, we knocked-down DDX21 via siRNA-mediated inhibition and assayed for 2′-O-methylation (2′-Ome) using site-directed cleavage of rRNA by RNaseH. Naive topological analysis suggested overlap in DDX21 binding to rRNA and 2′-Ome sites. We assayed resume of 2′-Ome in siRNA-treated cells using wild-type DDX21 ask well as a DDX21 mutation that lacks the ATPase domain. As expected, the DDX21 mutant failed to rescue the 2′-Ome defect.

The study highlighted a few reasonable points about these experiments and use of the pipeline: (1) Analyzing oft ignored classes of non-coding RNAs, such as snoRNAs, can reveal novel, testable hypothesis. (2) By producing data types that are easily comparable, the comparative analysis between different proteins or experiments is possible and useful. (3) CLIP-derived hypotheses should be tested using functional studies and experiment. In sum, it appeared that the assay and pipeline can reveal biological insights, which admit well to experimental validation.

5.5.2 Summary

The strategy we developed (termed FAST-iCLIP) incorporates a protocol that reduces experimental time by 50% with a computational pipeline that produces standardized data sets across protein coding, noncoding, and user-definable nonhuman transcriptomes. As sequencing continues to reveal novel noncoding RNA classes and further characterize microbial biodiversity, FAST-iCLIP can scale beyond the current human- and protein centric scope of CLIP study investigation.

Chapter 6

Infectious disease mechanism

Viruses are universally dependent upon their host cell. Despite the diverse functions that viruses encode for their propagation, they remain exquisitely dependent on the translational machinery of the host cell. No matter whether their genomes are RNA or DNA, and regardless of their mRNA production method, the goal remains the same: to ensure that cellular ribosomes are recruited to viral mRNAs [39].

Cellular mRNAs use cap-dependent translation, a process that involves interaction between initiation-factor proteins and the 7-methyl guanosine cap at the 5′ end of mRNA. This leads to 40S ribosome binding and scanning to the initiation codon, which is then followed by association with the 60S ribosomal subunit to form an active 80S ribosome that initiates translation of the protein (Figure 6.1).

An alternative pathway, called internal translation initiation, is a cap-independent mechanism of recruiting, positioning, and activating the eukaryotic protein-synthesis machinery driven by structured RNA sequences called internal ribosome entry sites (IRESs) that are located in the 5′ -untranslated region (UTR) of certain mRNAs [16]. Lacking a 5′ cap, many RNA viruses contain IRES that mediate cap-independent translation. In this process, the virus commandeers cellular ribosomes as well as translation factors and signalling pathways that control the host protein synthesis.

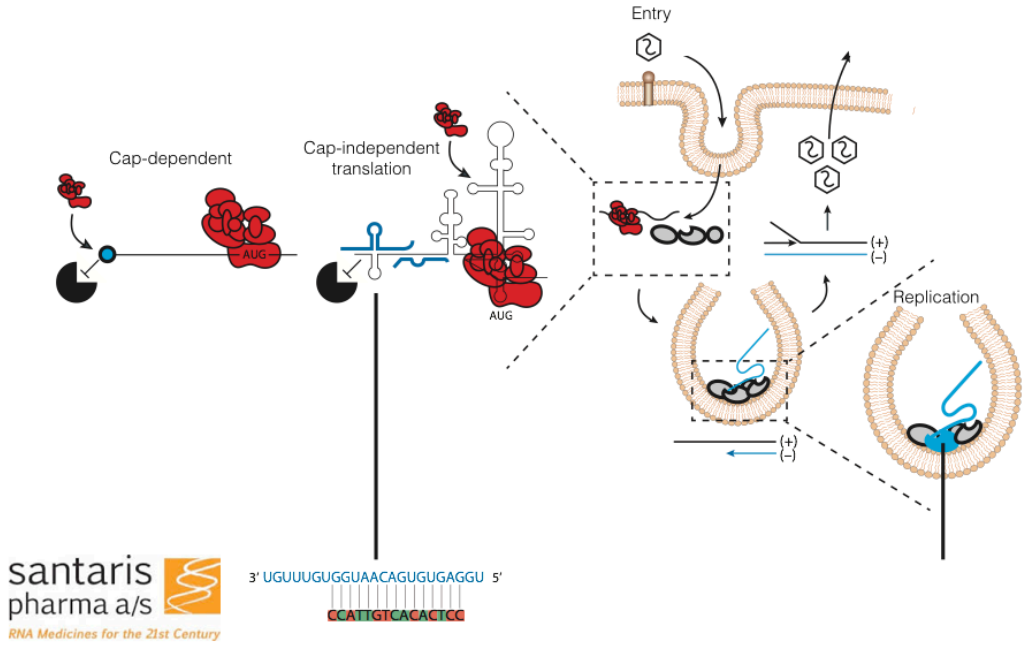

Figure 6.1: Different modes of translation.

6.1 HCV

IRES-mediated translation is well-studied in the context of HCV and the structure of the 5′ UTR region is well-understood [16]. Because of this, HCV is an excellent model system for extending the CLIP pipeline into infectious diseases. We approached this study by first identifying a protein that was (1) known to bind the HCV genome and (2) was critical for HCV replication, but (3) for which the mechanism of action remained unclear. We chose Poly-C binding protein 2 (PCBP2), a well-characterized RNA-binding protein with several studies linking it to HCV [17].

Though PCBP2 is required for HCV replication, the molecular details are poorly understood. Several studies focused on the HCV 5′ UTR have led to the suggestion that a complex between PCBP2 and SL1 of the 5′ UTR as well as an undefined region of the 3′ UTR of the viral RNA may be formed that facilitates viral circularization. Both SL1 and stem loop structures in the 3′ UTR of the viral genome are required

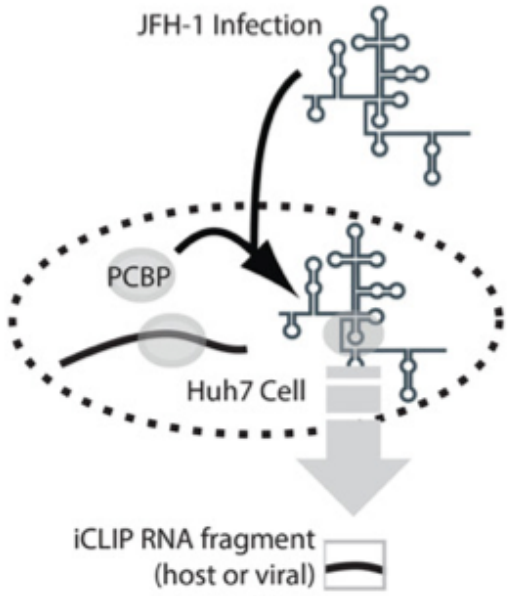

Figure 6.2: CLIP applied to HCV virus.

for viral RNA replication. In addition, the proximity of SL1 to the conserved miR-122 sites in the HCV genome suggests that PCBP2 may coordinate with miR-122 in protection of the uncapped 5′ end of the viral RNA from degradation and/or the switch between viral translation and RNA replication [17].

To elucidate the connection between PCBP2, translational regulation, and disease, we performed iCLIP in Huh-7 cells infected with the JFH-1 strain of Hepatitis C virus (HCV). We designed the pipeline so that it could easily be applied to viruses, and supplied the sequence of the JFH-1 genome as the mapping index (Figure 6.2). We generated coverage histograms of iCLIP RT stops across the HCV genome for two biological replicates, observing favorable concordance and global preference for binding U/C-rich regions of the genome ($r^2 = 0.93$) [12].

Consistent with prior studies, we observed a strong binding peak at SL1, but also detected PCBP2 occupancy that extends from SLI through the two miR-122 binding sites to the base of SL2. Surprisingly, we also detected strong binding around the translation start codon within SLVI of the internal ribosome entry site (IRES)

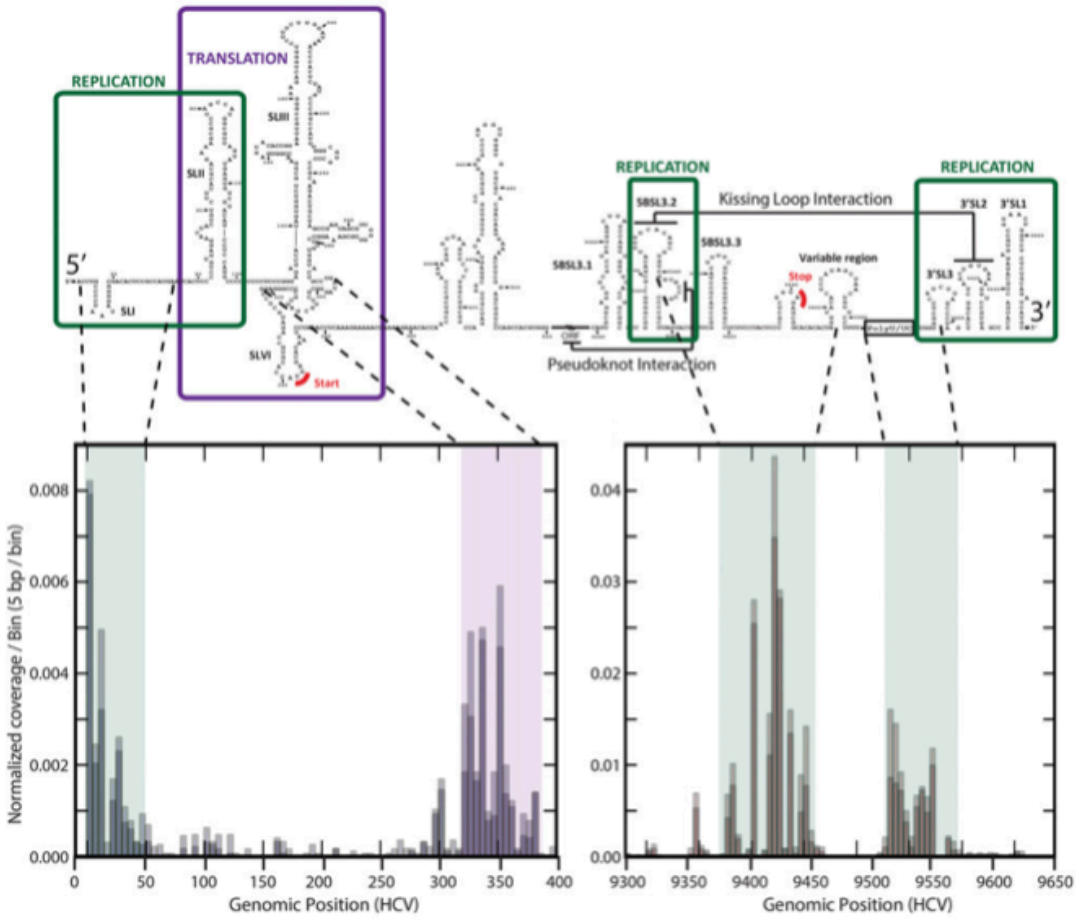

Figure 6.3: PCBP binding profile on the HCV genome.

(Figure 6.3). PCBP2′s interaction with viral 3′ UTR was significantly stronger than with the well-studied 5′ UTR. PCBP2 binding to the 3′ UTR occurred primarily in the single-stranded regions between stem-loops 5BSL3.2 and the variable region, a domain that includes the viral stop codon and that is implicated in both stimulation of translation and replication. Not surprisingly, PCBP2 also bound to the poly(U)/UC region of the viral genome, consistent with binding to single-stranded poly(U)/C regions. In addition to the UTRs, we observe multiple robust peaks of PCBP2 occupancy across the full viral gene body, which has never been reported.

Our application of FAST-iCLIP to HCV suggests that these regulatory functions

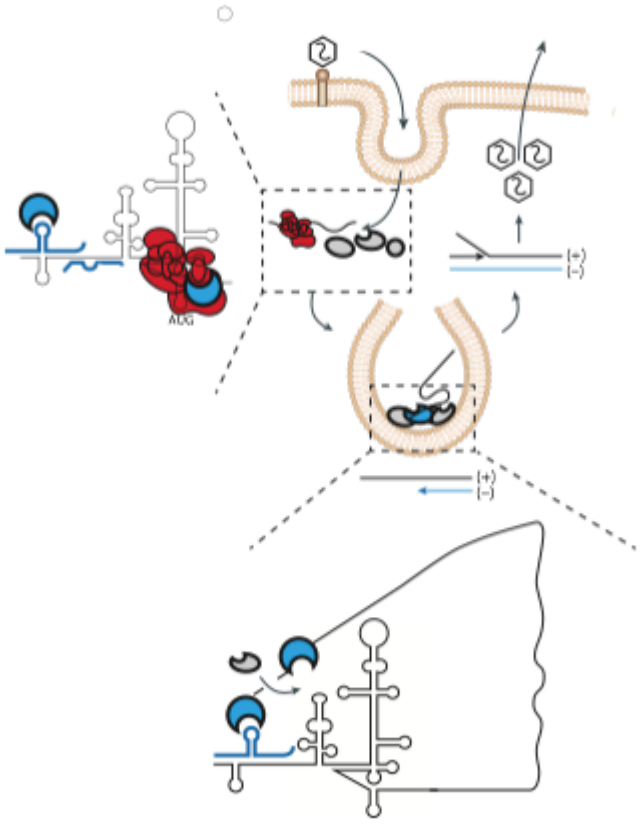

Figure 6.4: Models for PCBP and HCV infection.

of PCBP may be co-opted by the virus, as we also observe PCBP2 binding to the viral 5′ UTR, coding region, and 3′ UTR. We observe a peak of PCBP2 around the SL1/miR-122 binding site junction in the HCV genome, suggesting that PCBP2 may act in concert with miR-122 to restrict viral degradation from the 5′ UTR by cellular exonucleases such as Xrn2 [17]. PCBP2 also strongly bound to the translational start codon and the 3′ UTR of the HCV genome including the viral stop codon and conserved stem-loop structures required for viral RNA replication, a mode of binding that is topologically similar to that observed in poliovirus where it is well-known that PCBP2 plays a critical role in the viral life cycle [12].

In the context of a poliovirus infection, PCBP2 mediates cross-talk between the viral 5′ and 3′ UTRs in order to regulate the switch between viral translation and

Figure 6.5: CLIP applied to viral proteins.

RNA replication. Our data are consistent with a symmetrical role for PCBP2 in the context of HCV infection and overlays in vivo biophysical detail from prior reports showing PCBP2-mediates circularization of the HCV genome *in vitro* [12]. Thus, our application of FAST-iCLIP reveals a common binding topology of PCBP2 across the human transcriptome as well as the HCV genome. In both cases, a 3′ UTR bias is evident and suggests that PCBP2 regulatory functions may be co-opted by HCV.

6.2 Retroviruses

We have shown, for the first time, that CLIP-seq can be applied to pathogenic RNA viruses. The methods can reveal the molecular interactions between host proteins and the viral genome and is complimentary with high-throughput protein-protein interaction assays that have been applied in similar contexts [19]. Because these interactions are often required for viral replication and translation, information about binding topology can provide insight into mechanism [39] as well as reveal new host-viral interactions that guide development of new therapeutic strategies [20].

We next considered an alternative framing for viral CLIP experiments: the prior work assayed the binding topology of host protein on a viral genome. Could it be possible to investigate the binding of viral protein on host genes (Figure 6.5)?

Addressing this question requires a context in which viral proteins are expected to bind host mRNAs. For this, we turned to retroviruses, RNA viruses characterized by two unique, virally encoded enzymes (reverse transcriptase and integrase) in

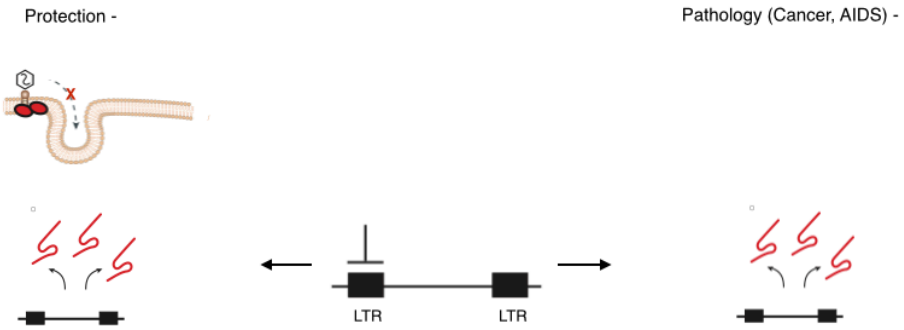

Figure 6.6: Two functional regimes for retroviruses.

their lifecycle. These enzymes permit the conversion of RNA into DNA, followed by the integration of viral DNA into the host genome, forming a DNA provirus [38]. Because retroviruses do not usually lyse their target cell, integration allows a long-term association between cell and virus. Numerous studies indicate that such events have occurred multiple times during the course of evolution. Such inherited proviruses are called endogenous retroviruses (ERVs) and they provide a fossil record of past retroviral infections dating back many millions of years.

Analyses of completed genome sequences have revealed that between four and ten per cent of vertebrate DNA is made up of retroviral remnants [38]. In general, the integrated 5 - 10 kb of sequence encodes for canonical retroviral gag, pol and env genes is flanked by two 300 - 1200 nucleotides bp terminal repeats (LTRs).

Each of these LTRs contains a promoter for RNA polymerase II and enhancers that are responsive to diverse conditions and signals. LTRs are bound by transcription factors, such as enable spatiotemporal control of proviral expression, though many ERV insertions are transcriptionally silenced in embryonic and adult tissues with repressive epigenetic marks [38]. Failure to maintain or contain these silencing marks would result in the reactivation of dormant ERV insertions. In contrast, some ERV-derived sequences that have been incorporated into the normal regulation of mammalian genes (promoters, enhancers or polyadenylation signals).

ERVs can be therefore thought of as either neutral, beneficial, or harmful. In the harmful case, retrovirses can be re-activated during immunosuppressive disorders,

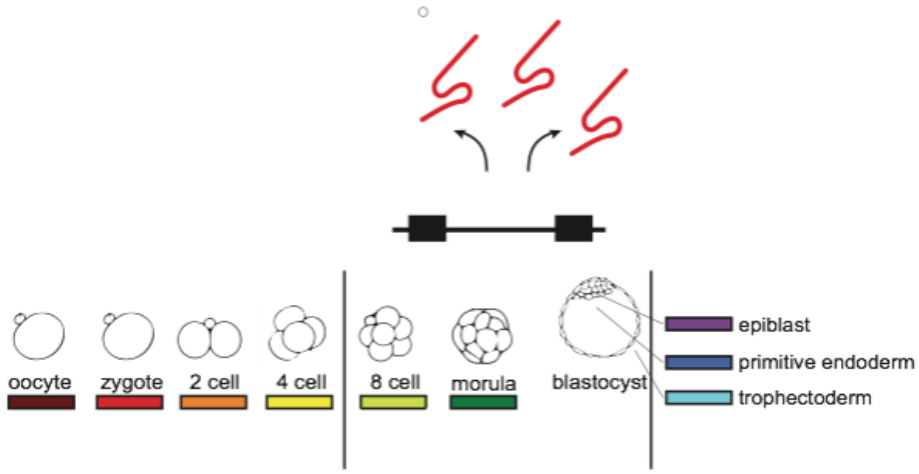

Figure 6.7: HERV-K is expressed in the naive human embryo.

such as AIDS or cancers. Expression of some retroviral Env proteins can modulate the host immune response. These ERV-encoded Env proteins possess immunosuppressive activities that can promote tumorigenesis, as indicated by several studies with mouse cancer models. In contrast, expression of ERV-encoded products can be beneficial. The Jaagsiekte sheep retrovirus (enJSRV) binds cellular receptors, blocking their ability to bind exogenous viral particles [37] (Figure 6.6).

Because retroviral proteins can interact with host factors, we examined the Human ERV-K (HERV-K) subfamily, which is apparently the most recent ERV wave to have entered our genome, perhaps producing insertions as recently as 150,000 years ago. HERV-K retained multiple copies of intact open reading frames (ORFs) that are silenced by the host with exception of certain pathological contexts, such as cancer or HIV [40]. For all human-specific and human-polymorphic HERV-K elements, silencing is mediated by by a specific LTR subgroup, LTR5HS.

Beyond obvious pathological contexts, recent single-cell RNA-seq has made it possible to examine HERV-K expression in naive embryos prior to blastocyst outgrowth. Analysis of repeat RNAs from single-cell RNA-seq dataset indicates that HERV-K is transcribed at the onset of embryonic genome activation (EGA), at the 8-cell stage, and detected in the embryo prior to EGA [44]. Human endogenous

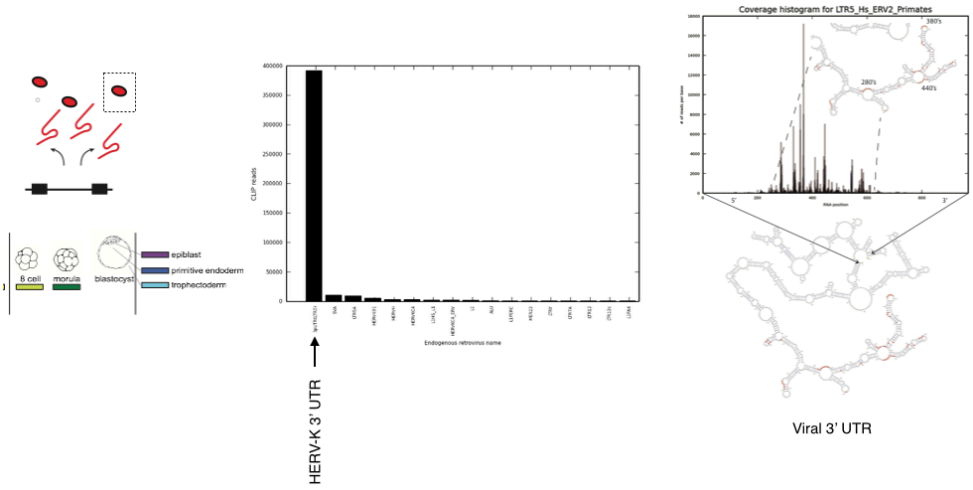

Figure 6.8: Rec binds the 3′ UTR of HERV-K in the naive human embryo.

retrovirus HERV-K and its regulatory element, LTR5HS, were induced after EGA in the 8-cell stage embryos and silenced during blastocyst outgrowth (Figure 6.7).

Prior studies suggest that retroviral products may be beneficial to the host by preventing exogenous viral infection [40]. We evaluated this by (1) asking whether retroviral products are made in the embyo, (2) examining the interaction between these products at the viral genome for evidence of bona-fide viral activation, and (3) assaying whether retroviral products can influence expression of anti-viral genes.

To evaluate whether HERV-K is functional in embryos, we assayed for the production of viral-like particles (VLPs), which can package viral RNA. We found that HERV-K VLPs indeed assemble in human blastocysts. We then used the CLIP pipeline to assay the viral protein Rec, which promotes nuclear export of viral RNAs.

Specifically, Rec is responsible for the export of unspliced and incompletely spliced viral RNAs to the cytosol. Rec binds as an oligomer to a responsive element (RcRE), which is located in the U3 region of the 3′ untranslated region of HERV-K transcripts [26]. If HERV-K is functional in human embryos, we expect this interaction to be observed. With this in mind, we performed CLIP in human embryonic carcinoma cells (hECCs) for Rec and evaluated the reads using a custom retroviral genome index that contained HERV-K as well as its UTR elements. We

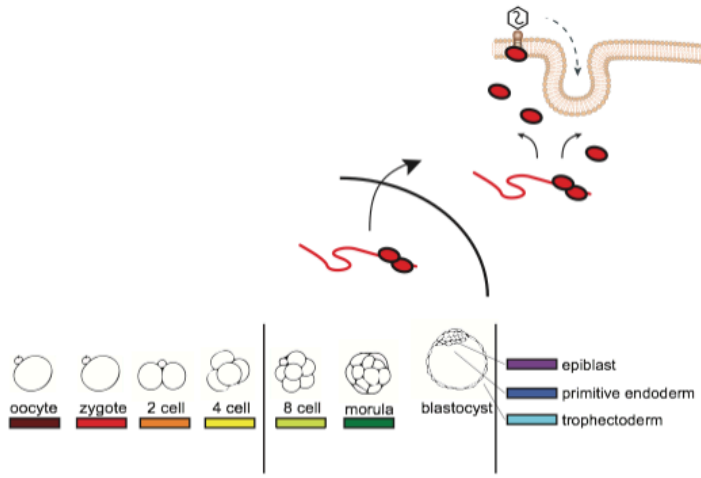

Figure 6.9: Retrovirals products in human embryo.

observed a strong signal in 3' UTR of the virus (Figure 6.7) in the region expected, the RcRE, which is indicative of viral activation in the embryo.

With this in mind, we asked whether Rec-mediated nuclear export of viral RNAs into the cytoplasm might lead to the upregulation of innate anti-viral response genes. We observed a striking upregulation of a viral restriction factor IFITM, with around 700 fold induction in human blastocysts as compared to pre-EGA (4-cell and 8-cell stage) embryos. IFTIM1, an interferon induced transmembrane protein, is thought to protect cells from viral infection at an early step by blocking endocytosis of virions into the cytosol. We also found that Rec overexpression is sufficient to increase IFITM1 levels in hECCs, the cell line in which we performed the CLIP.

Collectively, these results indicate that pre-implantation development proceeds in the presence of retroviral proteins, which may induce viral restriction factors that protect the embryo from exogenous infection. We used CLIP to evaluate the interaction network of a protein, Rec, that is critical in the HERV-K lifecycle. These results provide further evidence for activation of HERV-K and also reveal a cryptic network of interactions between Rec and host RNAs, which must be further characterized.

Chapter 7

Biophysical validation

7.1 The challenge with sequencing

The ability to generate large amounts of RNA-protein interaction data using CLIP-seq is useful [22]. But, how can be distinguish between bona-fide signal and artifacts products by non-specific molecular interactions? To date, our applications of CLIP relied on three approaches: (1) enforcing experimental replicates and merging reads from each [12], (2) standardized output data types that ease comparative analysis between datasets and make it easy to exclude genes with redundant, non-specific binding patters [12], and (3) experimental validation of hypotheses [6].

7.2 Highly multiplexed validation

We developed an additional method of validation, which takes advantage of microfluidic tools that enable highly multiplexed measurements of interaction affinity. Specifically, we adapted the MITOMI [13] microfluidic platform such that an entire RNA library (e.g., every RNA identified via CLIP) could be simultaneously synthesized and assayed for ligand-binding [27]. This technique, termed RNA-MITOMI, is thus complementary with sequencing-based methods for RNA interactomics (such as CLIP) because thousands of putative RNA-ligand interactions can be evaluated.

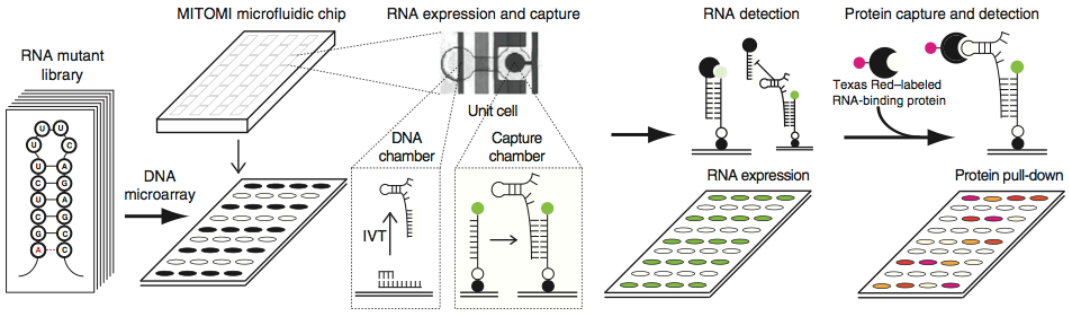

Figure 7.1: MITOMI assay design.

The MITOMI microfluidic chip used in this study has 640 microchambers, each with a volume ≈1 nl. We arrayed DNA oligos, which serve as transcription templates for the RNA library, and overlaid the MITOMI chip onto the array such that each spot was compartmentalized in a unique microchamber. Each chamber has a back-chamber, which houses the spotted DNA template, and a detection chamber in which the interaction between RNA and ligand is measured (Figure 7.1).

7.3 Stem loop binding protein

We applied this strategy to the well-characterized interaction between stem-loop binding protein (SLBP) and the 3′ histone mRNA stem-loop [28]. We measured relative SLBP binding across our library of single and double RNA mutants at 3 nM protein concentration (Figure 7.2). With this data, we identified three functional three regimes: 15 mutants had little effect on binding, 15 deleterious mutants were rescued by compensatory mutations that restore RNA structure, but three mutants could not be rescued by restoring structure, indicating sequence-based recognition at these positions. We used the data set to to distill sequence and structural requirements for stem-loop function (SLBP binding) at single-nucleotide resolution.

We validated the data in two ways. First, the features identified in the functional motif recapitulated the pattern of phylogenetic conservation of stem-loop sequences, suggesting that SLBP binding is the dominant selective constraint on

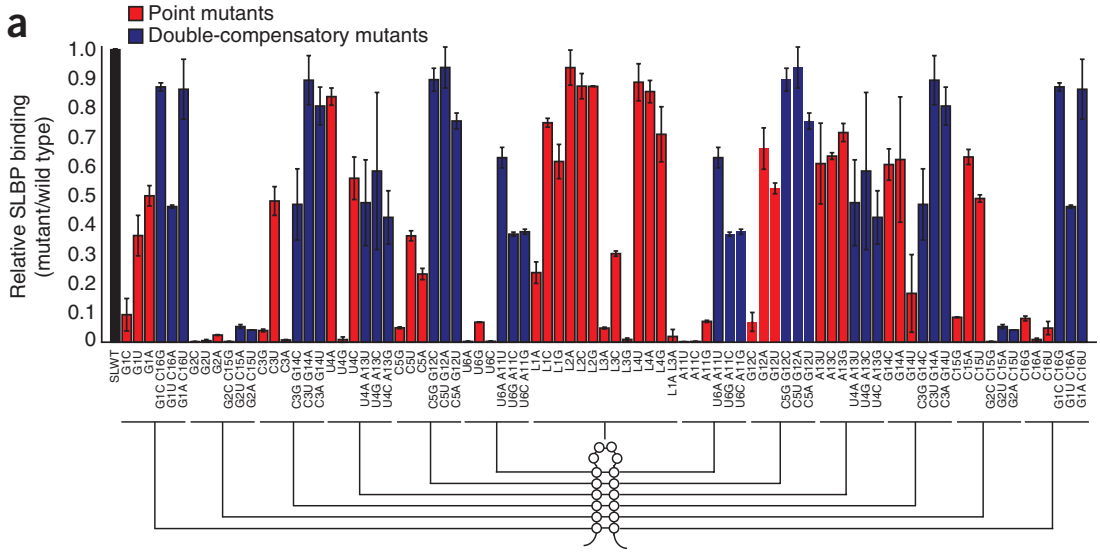

Figure 7.2: Multiplexed measurement of affinity.

the histone 3′ end. Residues (G2 and U9) and structural features (the U6 - A11 base pair) that are critical for SLBP binding are conserved from Tetrahymena sp. to humans, whereas we observed covariation for base pairs that have less stringent sequence-specificity [28]. Second, we tested nine point mutants with electrophoretic mobility shift analysis. The shift data agreed with the measurement of binding affinity obtained with RNA-MITOMI, confirming that measurements on the MITOMI platform can be recapitulated with conventional biochemical assays.

Our results show that SLBP recognizes much of the stem via RNA secondary structure rather than sequence (Figure 7.3). For nine of the 12 bases in the stem, function was not dependent on sequence, as individual base-pair substitutions at these positions were tolerated. In some cases, base substitutions were functionally neutral only if structure was preserved. In particular, Watson-Crick base pairing is required at the U6-A11 position at the top of the stem. Nucleotide-specific contacts are also important for SLBP binding. Consistent with prior studies, two residues in the loop (U7 and especially U9) as well as G2 in the stem are required for binding.

Extending the results from prior studies, using our panel of mutants we also identified noncanonical base pairs that are tolerated. Though the G2?C15 base pair

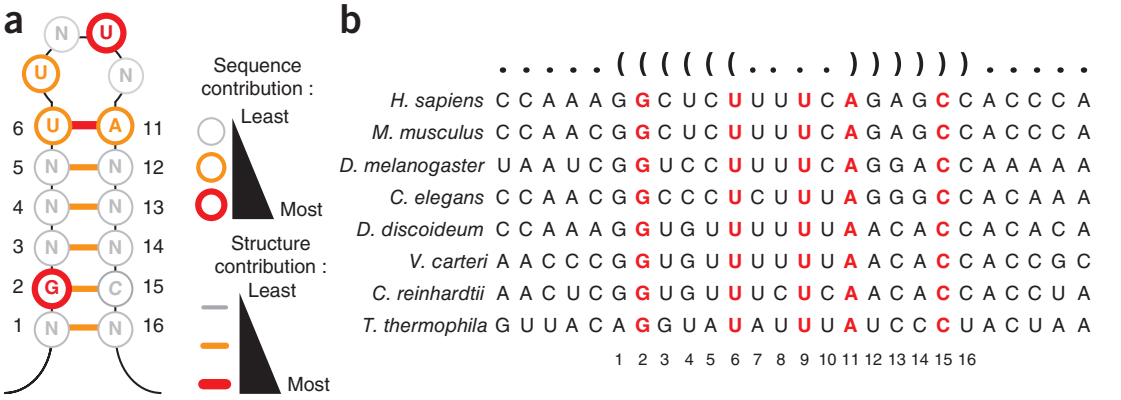

Figure 7.3: Function RNA motifs.

cannot be replaced by any of the three Watson-Crick base pairs, base substitutions in the position (C15) opposing G2 are tolerated, as mutations that allow wobble base-pairing (C15U) or prevent Watson-Crick base pairing (C15A) had nearly wild-type binding. Similar to the C15A mutation, C5A established a functionally tolerated G-A base pair. In addition, G12A and G1A mutations created a functionally tolerated A-C pair and G14C created a tolerated C-C. These noncanonical pairs can be deleterious if combined, suggesting limited tolerance to structure perturbation in the stem.

7.4 Summary

Highly multiplexed biophysical methods, such as MITOMI, can overcome the gap between high sequencing throughput and the conventionally low throughput of biophysical assays, such as gel-shift. Recently, these assays have be re-purposed on sequencing flow-cell [5], which suggests that read-out of CLIP-seq data could be followed directly by validation of biophysics on the same sequencing platform.

Chapter 8

Conclusions

8.1 The case of NGS and infectious disease

Infectious diseases have a profound impact on humankind, influencing the course of wars and the fate of nations. The traditional microbiology lab methods for detecting and identifying bacterial pathogens, notable culture, has limited resolution (e.g., provides no information about the genetics of the micro-organism) and scope (e.g., not every micro-organism can be cultured). By 1980, only 1800 validated bacterial species had been published. Yet, the advent of DNA-based methods - notably NGS - has rapidly accelerated micro-organism discovery (Figure 1.1).

The case for NGS in clinical microbiology is primarily driven by two points: (1) Conventional diagnostic testing for pathogens still fails to detect the causal agent in a significant percentage of cases [**?**] and (2) failure to accurately diagnose and treat infection in a timely fashion contributes to continued transmission and increased mortality in hospitalized patients. Furthermore, a flood of studies have shown that NGS is very useful for unbiased screening of rare or exotic pathogens, monitoring outbreaks or microbial resistance, and understanding broad compositional changes in microbial populations, such as a the human microbiome.

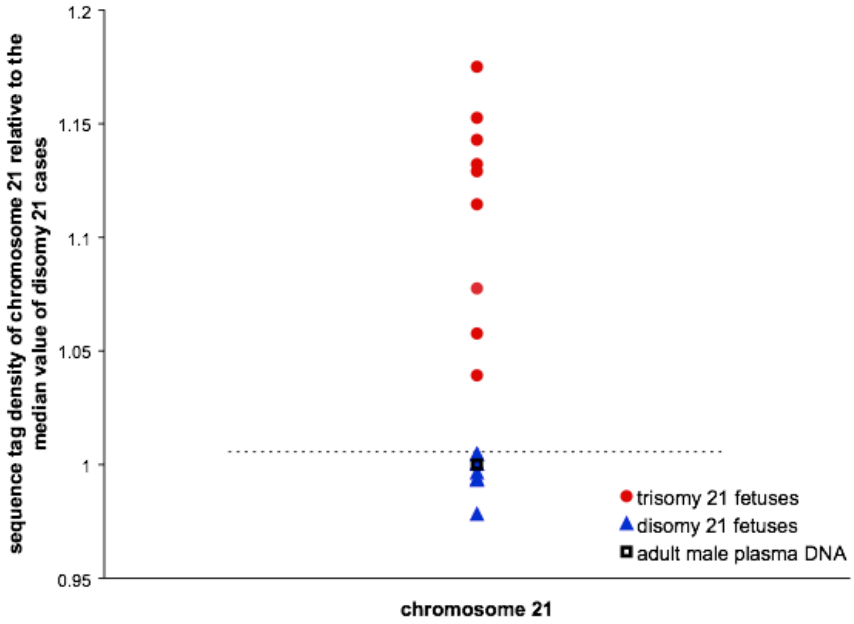

Figure 8.1: Molecular counting applied to chromosomes.

8.2 The cell-free DNA opportunity

A parallel thread of work presents an opportunity for NGS and infectious disease. This work is rooted in an observation made in 1947 by Mandel and Metais, who first note that human blood contains circulating cell-free DNA molecules. These fragments enter blood as the detritus of dead cells and circulate with short (15 minute) half-life as nucleosome-protected fragments. Methods of molecular counting (notably NGS) have taken advantage of this phenomenon. The greatest value has been to measure the proportion of foreign genomes within an individual.

The flagship application for NGS-enabled molecular counting was aneuploidy detection. Around 60000 invasive prenatal genetic diagnoses were carried out in 2009 in the United States, just over 1% of the 5.2 million women in the country who were pregnant that year [2]. Each test carries a slight risk of miscarriage, making

invasive testing particularly un-desirable. By counting chromosome derived cell-free DNA fragments using NGS, fetal aneuploidy is detectable, as counts mapping to the affected chromosomes (e.g., Chr21 for Down′ Syndrome) will be different than normal (e.g., elevated for chr21 trisomy in Down′ Syndrome) (Figure 8.1).

Because NGS-enabled molecular counting has been particularly effective for the detection of foreign genomes, it is natural to consider its application to infectious disease. In addition to the well-appreciated benefits of NGS-based testing for infectious disease, molecular counting of cell-free DNA would have additional benefits: (1) it may indicate infectious at any location within the body, rather than relying on correct selection of a particular fluid or biopsied sample, (2) it may replace or reduce the need for invasive testing in some clinical contexts, and (3) it may be useful as an indicator of microbial translocation from body sites into blood [36].

8.3 Quantifying micro-orgamisms in cell-free DNA

In order to build a diagnostic based upon molecular counting of pathogen-dervied cell-free DNA, these fragments must be sequenced, counted, and the results must be organized logically such that they are clinically informative. We built a pipeline capable of isolating micro-organism derived cell-free DNA fragments. Using this pipeline, we mined thousands of existing cell-free DNA database from multiple organ transplant cohorts at Stanford hospital. We then built a application that translates the raw data into easily interpretable statistics.

8.4 Molecular counting for pathogen diagnostics

Because infections are a major post-transplant complication, we obtained rich clinical testing data for many of the patients in the cohorts processed. We used this data in order to evaluate the clinical utility of our measurements. We found that molecular counting of infection-derived cell-free DNA fragment correlates favorably with conventional clinical test for numerous viruses, including CMV and Adenovirus.

We also identified favorable correlation for certain bacterial and fungal infections detected in various clinical samples, including urine and deep lung aspirate. In general, we found variable performance on bacteria and fungi. This is due to three issues: (1) We observe better performance on clinical specimens that have better coupling to blood (e.g., deep tissue aspirate and qPCR applied directly to plasma) relative to more peripheral fluids (e.g., skin or sputum). (2) We observe better performance on unique pathogens (e.g., CMV) relative than commensal human flora. While commensal flora can indeed cause pathogenic infection in certain circumstances, these micro-organisms can likely enter blood through various tissue sources and, in turn, have a high background signal. (3) We observe better performance for micro-organisms tested using specific molecular diagnostics, such as qPCR or MALDI-TOF. The performance is impressive, considering that this data was not enriched for non-human derived cell-free DNA. Infection derived cell-free DNA was not deeply sampled, as we only recovered hundreds of non-human derived reads from a sequencing depth of 30 million total reads for many of the samples. In spite of this, we have observed favorable clinical correlations on viruses, bacteria, and fungi, including cases of known deep-tissue infection. Furthermore, our results strongly highlight that unbiased NGS-based screening is a powerful way to detect many infections that escape clinical testing. These results will be further improved with biochemical strategies to deplete human-dervied DNA fragments.

8.5 Molecular counting for pathogen mechanism

While we have shown that NGS has great promise for infectious disease diagnostics, therapies are required to eradicate infections once they are identified. Most existing treatments target molecular networks that underpin the replication of infections agents. In turn, deeper understanding of these network has identified novel target and informed new treatments [**?**]. Human proteins are central to the propagation of pathogens, particularly viruses, which co-opt their host proteins for replication. RNA viruses and their association with human host proteins are particularly well-studied, as structural studies have provided great insight [16]. With this in mind,

we focused on RNA viruses and developed NGS-based pipeline (FAST-CLIP) that can be used to measure viral-protein interaction networks.

We first applied this to the interaction between human protein PCBP and the Hepatitis C virus. Though it was known that this protein is required for replication of the virus, our data precisely identified the genome-wide contacts between PCPB2 and HCV, indicating (1) contacts near the translation start site that indicate possible role in facilitating IRES-dependent ribosome loading, (2) contacts at both 5′ and 3′ UTRs indicating possible role for viral circularization during replication [17], and (3) a set of un-characterized interactions along the gene body of the virus.

We next applied the same pipeline to a different context, focusing instead on viral protein binding to host transcripts. We chose to examine the HERV-K endogenous retrovirus, as recent evidence suggested that this virus is unregulated in 8 day embryos. We performed CLIP on the retroviral protein Rec, showing that it is active in embryos and binds the HERV-K 3′ UTR in a well-studied viral motif that is required for nuclear export of the viral genome. Up-regulation of Rec induces antiviral host proteins, potentially protecting the embryo from exogenous viral infection. Rec also bind a large set of host transcripts, with functional consequences that are now being explored.

8.6 Summary and perspective

We showed that molecular counting of pathogen-derived cell-free DNA is a powerful diagnostic strategy. We built a pipeline for counting pathogen-derived cell-free molecules in human plasma and a web application presenting the resulting data. We applied these tools to thousands of clinical samples collected from hundreds of patients at Stanford hospital. We further processed thousands of clinical test records in order to show that this method can be broadly applied for non-invasive monitoring of viral, bacterial, and fungal infections in deep tissues. We further showed that unbiased pathogen monitoring using this technique has the potential to track many rare or un-expected infections that now escape hypothesis-centric clinical testing.

After demonstrating this new diagnostic application of NGS, we then showed

how NGS technology can be used to understand infectious disease mechanism. We developed a pipeline for counting of sequencing reads derived from RNA-protein interactions *in vivo*. For the first time, we showed that this method (CLIP-seq) can be applied to viruses that have infected human cells. We first used it to reveal novel interactions between the HCV genome and human protein PCBP2. In a follow-up study, we applied the method to HERV-K, an endogenous retroviruse. We showed for the first time that human embryo development occurs in the presence of retroviral products, which may protect the embroyo from exogenous infection while exerting regulatory function through interaction with human mRNAs.

We highlight three separate ways to validate the results from mechanistic CLIP-seq experiments, including comparative analysis, stringent replicate matching, and functional studies. To supplement these, we also developed a novel microfluidic strategy for highly multiplexed RNA-protein interaction measurements. This technique could be used to value all CLIP-seq tags from a given experiment in parallel, a throughput that far exceeds common biochemical assays. We applied this to a model RNA-protein interaction, recapitulating two decades of biochemical studies in a single experiment while also discovering novel features of the interaction.

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] D. W. Bianchi. Prenatal diagnostics: Fetal genes in mother's blood. *Nature*, 487(7407):304–305, 2012.

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

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

[5] J. D. Buenrostro, C. L. Araya, L. M. Chircus, C. J. Layton, H. Y. Chang, M. P. Snyder, and W. J. Greenleaf. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature Biotechnology*, 32(6):562–568, Apr. 2014.

[6] E. Calo, R. A. Flynn, L. Martin, R. C. Spitale, H. Y. Chang, and J. Wysocka. RNA helicase DDX21 coordinates transcription and ribosomal RNA processing. *Nature*, pages 1–17, Nov. 2014.

[7] E. D. Clercq. Antivirals and antiviral strategies. *Nature Reviews Microbiology*, 2(9):704–720, Sept. 2004.

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

[9] 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.

[10] 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.

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

[12] R. A. Flynn, L. Martin, R. C. Spitale, B. T. Do, S. M. Sagan, B. Zarnegar, K. Qu, P. A. Khavari, S. R. Quake, P. Sarnow, and H. Y. Chang. Dissecting noncoding and pathogen RNA–protein interactomes. *Rna-a Publication of the Rna Society*, 21(1):135–143, Dec. 2014.

[13] P. M. Fordyce, D. Gerber, D. Tran, J. Zheng, H. Li, J. L. DeRisi, and S. R. Quake. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotechnology*, 28(9):962–967, Aug. 2010.

[14] 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.

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

[16] C. S. Fraser and J. A. Doudna. Structural and mechanistic insights into hepatitis C viral translation initiation. *Nature Reviews Microbiology*, 5(1):29–38, Nov. 2006.

[17] A. Garcia-Sastre and M. J. Evans. miR-122 is more than a shield for the hepatitis C virus genome. *Proceedings of the National Academy of Sciences*, 110(5):1571–1572, Jan. 2013.

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

[19] S. Jäger, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, and K. Li. Global landscape of HIV-human protein complexes. *Nature*, 2011.

[20] C. L. Jopling. Modulation of Hepatitis C Virus RNA Abundance by a Liver-Specific MicroRNA. *Science*, 309(5740):1577–1581, Sept. 2005.

[21] M. Kleines, M. Häusler, A. Krüttgen, and S. Scheithauer. WU Polyomavirus (WUPyV): A Recently Detected Virus Causing Respiratory Disease? *Viruses*, 1(3):678–688, Dec. 2009.

[22] J. König, K. Zarnack, N. M. Luscombe, and J. Ule. Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, 13(2):77–83, 2012.

[23] 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.

[24] A. S. Lauring, J. Frydman, and R. Andino. The role of mutational robustness in RNA virus evolution. *Nature Reviews Microbiology*, pages 1–10, Mar. 2013.

[25] K. Lewis. Platforms for antibiotic discovery. *Nature Reviews Drug Discovery*, 12(5):371–387, Apr. 2013.

[26] R. Lo wer, J. Lo wer, and R. Kurth. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences.

[27] L. Martin, M. Meier, S. M. Lyons, R. V. Sit, W. F. Marzluff, S. R. Quake, and H. Y. Chang. Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. *Nature Methods*, 9(12):1192–1194, Nov. 2012.

[28] W. F. Marzluff, E. J. Wagner, and R. J. Duronio. Metabolism and regulation of canonical histone mRNAs: life without a poly (A) tail. *Nature Reviews Genetics*, 9(11):843–854, 2008.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

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

[35] 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.

[36] S. T. Sonis. The pathobiology of mucositis. *Nature Reviews Cancer*, 4(4):277–284, Apr. 2004.

[37] T. E. Spencer, M. Mura, C. A. Gray, P. J. Griebel, and M. Palmarini. Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. *Journal of Virology*, 77(1):749–753, 2003.

[38] J. P. Stoye. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews Microbiology*, 10(6):395–406, May 2012.

[39] D. Walsh and I. Mohr. Viral subversion of the host protein synthesis machinery. *Nature Reviews Microbiology*, 9(12):860–875, Oct. 2011.

[40] F. Wang-Johanning, A. R. Frost, B. Jian, L. Epp, D. W. Lu, and G. L. Johanning. Quantitation of HERV-K env gene expression and splicing in human breast cancer. 2003.

[41] J. A. Wilson and S. M. Sagan. Hepatitis C virus and human miR-122: insights from the bench to the clinic. *Current Opinion in Virology*, 7:11–18, Aug. 2014.

[42] 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.

[43] 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.

[44] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology*, pages 1–12, Aug. 2013.

[45] K. Zarnack, J. König, M. Tajnik, I. Martincorena, S. Eustermann, I. Stévant, A. Reyes, S. Anders, N. M. Luscombe, and J. Ule. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell*, 152(3):453–466, Jan. 2013.