METAGENOMIC METHODS IN IDIOPATHIC AND IMMUNE DISEASES

Erik L. Clarke

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Frederic D. Bushman, Ph.D., Professor of Microbiology

Graduate Group Chairperson

_____

Li-San Wang, Ph.D., Professor of Pathology and Laboratory Medicine

Dissertation Committee

Hongzhe Li, Ph.D., Professor of Biostatistics

Ronald G. Collman, M.D., Professor of Medicine

Elizabeth A. Grice, Ph.D., Assistant Professor of Dermatology

Luigi D. Notarangelo, M.D., Chief, Immune Deficiency Genetics Section, NIH

For my grandfather.

1935–2017

ACKNOWLEDGMENTS

This work would not have been possible without the mentorship and guidance of my advisor, Rick Bushman. From him, I have learned a tremendous amount about what makes a good scientist: the value of rigor and precision in both our methods and communications; of patience and understanding when working through mistakes (both yours and others); and of genuine enthusiasm for every level of the research process. I also want to thank Ron Collman, whose co-mentorship has been both inspiring and valuable in refining my research. Ron's confidence in my abilities and willingness to challenge my assumptions made me a better, more conscientious scientist. I am grateful to my thesis committee, Hongzhe Li, Elizabeth Grice, and Luigi Notarangelo, whose advice and guidance was essential for my development. I am also thankful to be part of the Genomics and Computational Biology graduate group, who pushed me to do more than I ever thought I could, and whose students and faculty have been good friends and mentors.

I am especially thankful for my fellow lab members over the past five years. Kyle Bittinger, Nirav Malani and Scott Sherrill-Mix are responsible for teaching me the foundations of R and reproducible research through their extraordinary patience and enthusiasm. Frances Male had the unenviable task of teaching me wet lab techniques, but despite this, she has been a talented collaborator and close friend. Young Hwang had the same thankless task as Frances, and without his mentorship and extraordinary help I could not have finished this dissertation. Aubrey Bailey and Chris Hoffmann were always sources of help and encouragement, both scientifically and socially, and made my time

here substantially more enjoyable. I also want to thank my fellow graduate students for their friendship and support: Arwa Abbas, Louis Taylor, Meagan Rubel, Christel Chehoud, Alexandra Bryson, and Sesh Sundararaman. I am also deeply thankful for everybody in the Bushman Lab, who together comprise one of the friendliest and most intelligent groups of scientists I know: Chris Nobles, Jesse Connell, John Everett, Kevin McKormick, Grant Eilers, Audrey Allen, Varun Aggarwala, Guangxiang Liang, Lyanna Kessler, Abigail Lauder, Jacob Leiby, Aoife Doto, Shantan Reddy, and John Gregg. I also want to express my gratitude for Laurie Zimmerman, my friend and guide through all of Penn's peculiarities. And I am thankful for the chance to work with many talented researchers at the PennCHOP Microbiome Program, including Kyle Bittinger, Ceylan Tanes, Chunyu Zhao, and Casey Hofstaedter.

I also want to thank the friends I made here outside the lab, including Shilpa Iyer, Ronnie Russell, Adam Hawkins, Stephanie Trimboli and Casey Hofstaedter. I could not have asked for a more entertaining and encouraging group of friends.

I want to thank my whole family, especially my parents, Leanne and Andy, whose unwavering love and support are the foundations I've built everything on, and my grandparents, whose inspiration and love made me the person I am today.

But most of all, I want to thank my husband Cory, my rock and best friend, who has never stopped believing in me even when I didn't believe in myself, who has been at my side through the best and worst times, and with whom I can do anything.

ABSTRACT

METAGENOMIC METHODS IN IDIOPATHIC AND IMMUNE DISEASES

Erik Clarke

Frederic D. Bushman

Microbial involvement in disease has been long-established, but only recently has metagenomics–the study of entire microbial communities–been leveraged in disease research. Powered by advancements in sequencing technologies, metagenomics has been used to better understand a wide variety of diseases, especially in the gastrointestinal tract and other areas with high microbial activity. However, the extreme sensitivity of modern sequencing and lack of established best practices yields high error rates when used in settings with low microbial involvement. To realize the potential of metagenomic sequencing in disease research, we need new high-precision experimental and statistical methods that reach clinical standards of confidence. In this thesis, I describe our search for microbial signatures in sarcoidosis, an idiopathic disease with suspected microbial involvement. Through the use of novel experimental and statistical methods, we were able to eliminate confounding environmental factors and identify the enrichment of Cladosporiaceae fungi in sarcoidosis. I next describe a computational method to recover pathogen genomes from a sample containing mostly host DNA without the use of over-sequencing or culturing. This method enables study of cryptic pathogen genomes and the ability to track genetic variants that may affect virulence or antibiotic resistance. Finally, I demonstrate the integration of metagenomic sequencing with immune repertoire sequencing in patients with severe combined immunodeficiency after gene therapy. This

study is the first to describe the changes in an immune-naïve microbiome that occur during the development of a new immune system. We see that the microbiome of these children shift from an abnormal state to one resembling healthy children in conjunction with their restored immunities. These studies lay out methods that improve the precision and utility of metagenomic sequencing for investigating idiopathic and immune disorders.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1. Introduction

1.1. A brief history of infectious disease

The birth of germ theory

The link between germs and disease seems obvious to us now, but it was actually a relatively recent (and contentious) development in human knowledge. For most of the last few thousand years, infectious disease was thought to be spread by miasma, or dirty air–a hypothesis that helped explain the transmissibility of disease in certain cases, but failed to translate into effective public health practices. As urbanization spread and diseases like cholera became more prevalent, water- or sewage-based transmission patterns played a large role in disease outbreaks. In a now-famous cholera outbreak in London, a physician named John Snow tracked each disease incident and concluded from their distribution that they originated from a public water pump. He correctly hypothesized that the water was contaminated and suggested the agent of contamination (what he termed "cholera poison") had a cellular structure and was capable of reproducing (Snow, 1856).

This hypothesis was part of an emerging "germ theory" of disease based on the groundbreaking work by Ignaz Semmelweis, Louis Pasteur and others that suggested that microorganisms, rather than miasma, were the source of infections and infectious disease. Joseph Lister's adoption of this theory and promotion of antiseptic practices in surgery (and the consequent reduction in mortality) showed germ theory's immediate

applicability to healthcare, but it was not until the work of Dr. Robert Koch that a conclusive link between a microbe and a human disease was first established.

Establishing the germ-disease link

Koch was the first to demonstrate the causative link between anthrax and *Bacillus anthracis* spores by purifying the bacteria from an affected sheep and using it to infect mice, who developed anthrax symptoms (Koch & Carter, 1987). Similar experiments using culture and reinfection lead him to uncover the cause of cholera (*Vibrio cholerae*) and tuberculosis (*Mycobacterium tuberculosis*) and conclusively demonstrate a link between microbes and infectious disease (Koch & Carter, 1987). This work also set the foundation for modern culture-based microbiology by improving culture and staining techniques, but most importantly it led to the development of what are now known as Koch's postulates.

Koch's postulates are guidelines for establishing a causal link between an organism and a disease (Koch, 1876). They require a) the presence of the organism in all affected subjects and absence in unaffected subjects; b) the organism must be isolated and cultured from an affected subject; c) when introduced to an unaffected subject, the isolated organism should trigger the disease; and d) the identical organism must be cultured from the newly-diseased subject.

Successes of Koch's postulates

The establishment of the germ-disease link and Koch's postulates led to a revolution in medical microbiology as the causes of some of humanity's most troublesome diseases

were uncovered. In this "Golden Age" of microbiology (Blevins & Bronze, 2010), bacteria causing diphtheria (1883), tetanus (1884), pneumonia (1886), plague (1894), and dysentery (1898), as well as many others (Table 1-1), were identified using the principles behind Koch's postulates. These discoveries were essential in tracing disease vectors– often involving fleas, rodents, or sewage–and establishing key elements of modern public health practices, sanitation, vaccines, and epidemiology. Later, the advent of antibiotics would provide the means to treat, rather than simply prevent, these diseases.

Year	*Disease*	*Organism*	*Discoverer*
1877	Anthrax	*Bacillus anthracis*	Koch
1878	Suppuration	*Staphylococcus*	Koch
1879	Gonorrhea	*Neisseria gonorrhoeae*	Neisser
1880	Typhoid fever	*Salmonella typhi*	Eberth
1881	Suppuration	*Streptococcus*	Ogston
1882	Tuberculosis	*Mycobacterium tuberculosis*	Koch
1883	Cholera	*Vibrio cholerae*	Koch
1883	Diphtheria	*Corynebacterium diphtheriae*	Klebs, Loeffler
1884	Tetanus	*Clostridium tetani*	Nicholaier
1885	Diarrhea	*Escherichia coli*	Escherich
1886	Pneumonia	*Streptococcus pneumoniae*	Fraenkel
1887	Meningitis	*Neisseria meningitidis*	Weischselbaum
1888	Food poisoning	*Salmonella enteriditidis*	Gaertner
1892	Gas gangrene	*Clostridium perfringens*	Welch
1894	Plague	*Yersinia pestis*	Kitasato, Yersin
1896	Botulism	*Clostridium botulinum*	van Ermengem
1898	Dysentery	*Shigella dysenteriae*	Shiga
1900	Paratyphoid	*Salmonella paratyphi*	Schottmüller
1903	Syphilis	*Treponema pallidum*	Schaudinn, Hoffmann
1906	Whooping cough	*Bordatella pertussis*	Bordet, Gengou

Table 1-1 – Discoveries from the 'golden age' of bacteriology.
Researchers who discovered the bacterial pathogens behind mankind's most troubling diseases during the 'golden age' of bacteriology. Adapted from Blevins and Bronze (2010).

1.2. Sequencing and its use in infectious disease

Reaching the limits of culture-based investigations

While the principles behind Koch's postulates remain the gold standard for identifying pathogens, it became quickly evident that they were not universally applicable. Koch himself noticed the presence of *M. tuberculosis* and *V. cholerae* in subjects who did not demonstrate symptoms, violating the first postulate (and establishing the concept of an asymptomatic carrier). Another persistent complication is the fact that many organisms are not able to be grown in pure culture, and so the experimental design suggested by Koch (culture and experimental infection in an animal model) is inapplicable.

There are any number of reasons why a pathogen might be unculturable. For instance, viruses require specific cellular machinery to reproduce, so culturing on cell-free medium is impossible. In other cases, the parasite may require coinfection with another organism for viability (Hepatitis D, for instance (Makino et al., 1987)). This lack of cultivability does not preclude these organisms from being pathogenic, however, so rigid adherence to Koch's postulates would preclude investigators from understanding the diseases caused by these types of microbes. Adaptation around these limitations lead to the discovery of the viral causes of yellow fever (Sellards & Hindle, 1928), foot-and-mouth disease (Loeffier & Frosch, 1897), and polio (Landsteiner K, 1909), among many others. These advancements demonstrated a growing understanding of the range and diversity of pathogens.

One approach toward the investigation of uncultivable pathogens involves the use of nucleic acid sequencing. Sequence-based methods provide researchers with ways to identify pathogens regardless of their cultivability. Approaches such as PCR for specific marker sequences in target organisms are exquisitely sensitive and do not depend on the viability of the target organism (Josephson, Gerba, & Pepper, 1993). More complex methods, including whole-DNA shotgun sequencing, are able to recover the entire genome of a microbe (Anderson, 1981)–or even the genomes of all the microorganisms in a sample (Segata et al., 2013). These methods have revolutionized microbiology because they have allowed us to understand more fully the microbial world outside the petri dish. They have also advanced infectious disease research because their sensitivity enables us to uncover broad spectrums of pathogens that are not able to be cultured or otherwise identified.

To formalize the use of sequencing methods in infectious disease research, Fredericks and Relman published an update to Koch's postulates (Fredricks & Relman, 1996). The new guidelines were specifically crafted to avoid the requirement of cultivability in suspected microbes, as by the time that review was published, countless diseases had been linked to microbes that cannot be grown in pure culture, such as Whipple's disease (the *Tropheryma whippelii* bacterium, (Relman, Schmidt, MacDermott, & Falkow, 1992)) and Kaposi's sarcoma (human herpesvirus 8, (Huichen Feng et al., 2007)). Notable differences between the Koch's and Relman's postulates, besides the de-emphasis on cultivability, include reframing the presence or absence of the organism as the enrichment or depletion of nucleic acids from that organism; correlation between

sequence- and tissue-based methods such as staining or *in situ* hybridization; and identification of plausible methods of virulence (e.g. from related species or identified virulence factors). The Relman guidelines are significantly less rigid in their formulation than Koch's postulates as well. For instance, the requirement of enrichment of nucleotide signals in affected tissues, rather than absolute presence/absence, reflects the understanding that many factors unrelated to the disease state may affect the appearance of microbial DNA, including environmental factors, presence as commensals, and reagent contamination. In total, Fredricks and Relman's guidelines are given as flexible measures that should be assessed holistically, much as Koch's postulates had to be later interpreted. The violation of multiple guidelines is possible even in established causal relationships and the authors emphasize that none of the rules are dogmatic.

Successes using sequencing in etiologic agent detection

Despite the incumbent difficulties in establishing microbe-disease links via sequencing, there have been a number of notable successes. In 1992, David Relman *et al.* established that Whipple's disease was caused by a previously-unknown microbe by recovering 16S ribosomal RNA sequences from five patients with Whipple's disease (Relman et al., 1992). The organism, termed *Tropheryma whippelii*, had been seen in affected patients via microscopy since the first description of the disease but had resisted all culture efforts. Its association with the disease was thus well-known, and the target sequence was isolated from all five patients and none of the ten healthy controls. This long-term association of the cryptic microbe and the disease thus simplified the challenge of establishing causality. While the precise virulence mechanisms were not described in the

1992 paper, they noted that *T. whippelii* was related to other *Actinomycetes* bacteria including pathogenic mycobacteria, and potentially shared disease-causing characteristics.

In 2008, Feng *et al.* (H. Feng, Shuda, Chang, & Moore, 2008) identified the virus that caused Merkel cell carcinoma (MCC) through the use of digital transcriptome subtraction (Huichen Feng et al., 2007). This technique uses a reverse transcription step to convert mRNA into cDNA prior to sequencing, and subsequently removes all human-related cDNA after sequencing. The remaining sequences are checked for homology to potential organisms of interest: in this case, they already suspected a viral etiology based on similarities between MCC and Kaposi's sarcoma. The presence of a transcript with similarities to existing polyomavirus T antigen allowed the researchers to eventually uncover the complete viral genome using primer walking; this genome was then used to check transcripts from other MCC libraries and to design PCR assays to check other MCC samples. The rates of detection for the new polyomavirus (termed Merkel cell polyomavirus) were 80% in MCC tissue and 8% in healthy; this represents a significant shift from the strict presence/absence suggested by Koch's postulates and a concordance with Fredricks and Relman's more flexible guidelines.

More recently, modern metagenomic sequencing was used to identify the pathogen in a case of neuroleptospirosis that had evaded diagnosis via normal methods (Wilson et al., 2014). A patient that presented with severe but ambiguous symptoms that were initially suspected to be neurosarcoidosis after no pathogen was detected. Cerebrospinal fluid

(CSF) and serum was then shotgun sequenced and analyzed with SURPI, a rapid pathogen detection pipeline (Naccache et al., 2014). SURPI detected traces of Leptospiraceae bacteria in the CSF but not in the serum sample with coverage across 3.8% of a leptospira genome. On this basis, the administering physicians decided to treat the patient for neuroleptospirosis and the symptoms resolved. Leptospiraceae is normally detected through a serological test for immune response, but due to peculiarities in the patient's condition (severe combined immunodeficiency, immunoglobulin supplementation, etc), the serological challenge was negative. In many ways, the success in identifying the causative agent in this study was unusually fortuitous. Leptospira are uncommon enough in a hospital setting not to be likely environmental contaminants and pathological enough in nature to be a convincing agent. For diseases where the etiologic agent is uncharacterized, SURPI is not a valid option because it references existing databases- a flaw intrinsic to most of the current methods of sequence classification. If the organism was commonly commensal and pathogenic only by merit of a compromised immune system, it would also be difficult to discern from background levels of that taxa.

Isolating microbial species for genomic analysis

Identifying the etiologic agents in disease is rarely the final step in treatment and diagnosis. Virulence and pathogenicity of an agent is often linked to species- or strain-specific variations that may be difficult to uncover using normal phylogenetic markers such as the 16S or ITS rRNA gene sequences (a method explained in more detail below). To fully understand the pathogenesis of an organism, the complete genome is desirable so that virulence mechanisms can be determined. However, genomes for most microbial

species are still missing or incomplete (Aggarwala, Liang, & Bushman, 2017; Brown et al., 2015; Hug et al., 2016).

Traditionally the means to get a precise species- or strain-level genome is to isolate and culture the organism and sequence the isolate, but as described above, such culture-based approaches are frequently not applicable (Amann et al., 1990; Ghazanfar, Azim, & Ghazanfar, 2010; Schmeisser, Steele, & Streit, 2007). In the examples outlined previously, few of the etiologic agents were cultivable in pure culture. One way around this is to sequence the sample with sufficient depth to achieve complete coverage of the target genome (along with a large amount of background DNA) (Forde & O'Toole, 2013; Mardis, 2008). In cases where the genome is especially small, including viruses, enrichment using PCR techniques are viable options (Minot et al., 2013), but these techniques are difficult to scale up to normal prokaryotic and eukaryotic genomes.

There have been significant improvements in the ability of shotgun metagenomic sequencing to identify strain-level variations in microbial communities (Alneberg et al., 2014; Olm, Brown, Brooks, & Banfield, 2017; Scholz et al., 2016), but these efforts all face similar pitfalls. For some methods, the complete strain-level genome must be captured and added to the relevant databases. For instance, programs like StrainPhlAn (Truong, Tett, Pasolli, Huttenhower, & Segata, 2017) and PanPhlAn (Scholz et al., 2016) use strain-level reference genomes or species-wide "pangenomes" to profile metagenomic communities for those known strains. Other approaches that avoid the need for reference strains, such as CONCOCT (Alneberg et al., 2014) and MetaBAT (Kang,

Froula, Egan, & Wang, 2015), use pooled metagenomic libraries to construct strain-level genomes, but face challenges when the organism of interest is especially rare in the sample and are prone to chimeric misassembly of strain genomes.

Methods such as selective whole-genome amplification (SWGA) (Leichty & Brisson, 2014) allow the enrichment of a target organism in a sample by exploiting differences in certain sequence motifs along the target and host DNA. By using primers targeted to sequence motifs more frequently appearing in the target than the background, and a highly-processive polymerase such as phi29, a researcher can preferentially amplify the target genome above the background genome. The resulting product is directly sequenced, but the resulting depth of sequencing required to reach sufficient genome coverage is substantially reduced. In this way, SWGA makes it easier to get precise variant-level information about an organism and allows the sequence variants that underlie its virulence or pathogenicity to be identified. Implementing SWGA for arbitrary genomes is not trivial, however, as the method described in the original paper for the selection of effective primer sets is complex and error-prone.

1.3. Metagenomics and the role of the microbiome in disease

The human microbiome

Infectious disease research focuses on the single causative agents of a disease: isolating and understanding particular pathogens and how to combat them. But it has been known for some time that the body is home to a vibrant community of microbial life. In 1673, Antonie Philips van Leeuwenhoek, an acclaimed microscopist, described seeing a variety

of tiny “animalcules” in samples of pond water and wood–making him the first human to see a bacteria. In 1861, Joseph Leidy described a flourishing community of microbes inside the guts of many animals. His work “A flora and fauna within living animals,” is one of the first descriptions of the microbiome (Leidy, 1861).

We know now that humans share our bodies with trillions of commensal microorganisms, most of them bacteria and bacteriophage. The exact makeup of these communities vary by body site and functional characteristics (The Human Microbiome Project, 2012). They also play key roles in maintaining normal health. For instance, microbes in the gut facilitate nutrient absorption and digestion (Shreiner, Kao, & Young, 2015), help regulate the immune system (Arpaia et al., 2013), and may even affect mental state (Rieder, Wisniewski, Alderman, & Campbell, 2017). Microbes on the skin feed off of shedding skin cells and help prevent infections (SanMiguel, Meisel, Horwinski, Zheng, & Grice, 2017), and microbes in the genital tract can hinder the acquisition of sexually-transmitted diseases, such as HIV (Buve, Jespers, Crucitti, & Fichorova, 2014).

The microbiome’s role in maintaining health is also supported by evidence linking disordered communities to diseases including inflammatory bowel diseases (Gevers et al., 2014), obesity (Le Chatelier et al., 2013), cardiovascular disease (Z. Wang et al., 2011), as well as lung and skin disorders (Charlson et al., 2010; Hannigan, Pulos, Grice, & Mehta, 2015; Kalan et al., 2016; Young et al., 2015). Thus, it is essential to understand exactly what makes a microbiome “healthy”, what functions they perform in their respective body sites, and how to alter their composition to prevent or treat disease.

The advent of metagenomic sequencing

Traditional culture-based techniques are inefficient for metagenomic studies both because of the relatively small fraction of microbes that are cultivable, and because of the slow, low-throughput nature of culturing. In addition, phylogenetic characterization using morphology rather than genetic sequence leads to significant issues: some microbes may be morphologically identical but separate species, while other microbes may have wildly different appearances based upon factors such as environment or lifecycle stage– some fungi belong to two different families depending on their sexual stage due to exactly this problem (Underhill & Iliev, 2014).

The concept of metagenomic sequencing is basically the adaptation of modern sequencing technologies to perform surveys of all the microbes present in a sample. Generally, metagenomic sequencing falls into two approaches: tagged marker sequencing or whole-genome shotgun sequencing. Tagged marker sequencing uses the amplification of conserved genetic regions in the microbial targets of interest to gather a picture of the microbial community in a sample. Examples of this include 16S rRNA sequencing for bacteria (Weisburg, Barns, Pelletier, & Lane, 1991), and internal transcribed spacer (ITS) rRNA sequencing for fungi (Schoch et al., 2012). These regions, or "markers," contain hypervariable loci flanked by conserved sequences; in bacterial 16S rRNA these are regions V1-V9, and in eukaryotic ITS rRNA they are ITS1 and ITS2. These hypervariable loci are under significantly less selective pressure than their surroundings, and thus accumulate mutations at a higher rate (Gray, Sankoff, & Cedergren, 1984). The similarity of these regions between two taxa can therefore be used as a proxy for how

related the taxa are to each other. This also enables the cataloging of these regions into databases, so that microbes can be linked to their respective version of each marker. Thus, the amplification and sequencing of these marker regions, followed by similarity searches in databases like GreenGenes for 16S (DeSantis et al., 2006) and UNITE for fungal ITS (Koljalg et al., 2005), results in a picture of the bacterial or fungal communities in a sample. Tagged sequencing is generally low-cost and computationally straightforward. However, the results are limited to the microbial kingdom of choice, and even the selection of the variable region in the marker sequence can influence how sensitive the assay is at retrieving certain families of microbes (Meisel et al., 2016).

The alternative approach to tagged sequencing is known as whole-genome shotgun sequencing. In this approach, the total nucleic acids in a sample are fragmented and sequenced without an intermediate amplification step, and the sequencing reads are matched to a database of microbial sequences (Quince, Walker, Simpson, Loman, & Segata, 2017). The benefits of this approach are that it is markedly less biased as it can capture DNA from any organism in the sample, and that it allows partial reconstruction of the genomes of the organisms, enabling functional characterization. Shotgun sequencing thus provides a more complete picture of the microbiome, but suffers from both high costs and analytic difficulty. In many shotgun sequencing experiments, a large fraction of the reads cannot be assigned to an organism due to the true originator of the read being absent from databases. And while the sequencing itself is relatively less biased, the databases themselves are not– some microbial orders are much better characterized than others (e.g. bacteria versus viruses or fungi).

As well as helping characterize the function and dysfunction of our microbiomes, metagenomic sequencing provides a key advancement in germ theory and our ability to associate microbes with disease. Sequencing the total DNA from healthy and diseased subjects enables detection of differentially-abundant microbial signatures and potentially illuminates cryptic agents in diseases that had evaded culture-based detection. Furthermore, the ability to gather the actual genetic sequences from these microbes provides the means to understand their pathogenic capacity. Genes that influence virulence, toxicity and resistance to antimicrobial compounds can be recovered from the metagenome and may one day form the basis of rapid sequencing-based clinical tests.

Successes linking the microbiome to disease state

Because the gut is the most well-characterized component of the human microbiome, many of the disease links we've associated with changes in the microbiome have to do with intestinal disorders. In particular, inflammatory bowel diseases such as Crohn's disease have been closely linked to dysbiosis in the gut (Huttenhower et al., 2014; Lewis et al., 2015). Crohn's disease is a complex disorder involving immune-mediated inflammation of the gut, especially the proximal colon and ileum. Instead of a single causative bacterium, researchers have found that structural changes, such as the loss of certain classes of bacteria and outgrowths of others, can be predictors of Crohn's disease onset (Gevers et al., 2014; Haberman et al., 2014). Antibiotic usage and a decrease in bacterial diversity were also associated with disease severity (Lewis et al., 2015). Correction of dysbiosis through the use of fecal microbiome transplantation has shown promising results for Crohn's disease (Ruben J Colman, 2014).

Perhaps the most classic case of microbiome dysbiosis correlating with disease is in the case of *Clostridium difficile* infection. *C. difficile* is a commensal microbe in infants but rarely present in asymptomatic adults (Rousseau et al., 2012). Instead, infection is usually triggered by exposure to antibiotics, especially long-term use in chronic care facilities (Britton & Young, 2014). The alteration of the endogenous flora through antibiotics seems to provide an avenue for *C. difficile* colonization, and indicates that a "healthy" microbiome provides resistance to this pathogen. The exact mechanism by which the microbiome changes from being protective to permissive seems to be related to the production of bile salts and secondary bile metabolites (Britton & Young, 2014) (Britton 2014). Certain bile acids trigger *C. difficile* spore germination, and their relative availability in the microbiome can be increased through the use of antibiotics. Thus, in the case of *C. difficile,* we observe a traditional infectious disease that is influenced strongly by the characteristics of the host gut microbiome.

The use of the dysbiosis as a disease marker has also been demonstrated in preterm infants for necrotizing enterocolitis (NEC) (Pammi et al., 2017). In a systematic meta-analysis, the authors found that increased levels of Proteobacteria and loss of Bacteroidetes and Firmicutes characterized a dysbiotic state that was a precursor to NEC. The correlation between gut bacteria and disease state indicate that microbiome monitoring and correction through the use of probiotics may be effective treatment for an otherwise difficult-to-treat disease.

Beyond disorders involving the gastrointestinal track, dysbiosis in the microbiome has also been linked to cardiovascular disease and mental health. In one study, Z. Wang et al. (2011) showed a link between the microbiome and production of phosphatidylcholine, a dietary metabolite correlated with heart disease. When the researchers dampened the abilities of the microbiome to produce this metabolite through the use of antibiotics, they saw a resulting reduction in disease rates, confirming the link between the metabolomic function of the microbiome and the disease. Recent studies have also linked the microbiome to mental health, mostly in animal models. For instance, in mice with a murine version of autism, researchers found they had abnormal microbiomes, and that addition of certain beneficial bacteria into their gut lead to less autistic-type behaviors (Hsiao et al., 2013).

The microbiome and the immune system

The links between microbiome structure and disease state often involve the immune system in some way (e.g. inflammation in IBD-type disorders). This is likely because the microbiome has been shown to have a regulatory and supplemental effect on the immune system itself (Arpaia et al., 2013; Atarashi et al., 2013; Kamada & Núñez, 2014). For instance, Arpaia et al. (2013) demonstrated that metabolites produced by the microbiome influence the development of pro- and anti-inflammatory regulatory T (Treg) cells. As these Treg cells are kept in balance in healthy individuals, it suggests that there is a homeostatic relationship governing the microbiome and immune system that can become disrupted in cases of dysbiosis.

The importance of the gut microbiota in maintaining immune homeostasis is in fact well documented (Guarner & Malagelada, 2003; Renz, Brandtzaeg, & Hornef, 2011; Walker, 2013). For instance, hypersensitivity of the immune system to food-based allergens appears to be moderated by the gut flora (Guarner & Malagelada, 2003). This effect is determined by the co-development of the immune system with bacterial colonization of the gut in early life and takes place on the mucosal interfaces of the intestine and respiratory tract (Renz et al., 2011). In general, abnormal colonization or immune conditions early in life can lead to dysbiosis, allergy, or more dangerous immune disorders (Walker, 2013).

Pitfalls in metagenomic sequencing

Metagenomic sequencing thus has demonstrable clinical value in diagnosing and understanding a variety of diseases. However, there remain significant experimental and technological challenges associated with it that hinder clinical adoption. Perhaps most critical from a diagnostic or etiologic standpoint is modern sequencing's susceptibility to false positives. This is partially due to the extreme sensitivity of sequencers and library preparation methods (Chin, da Silva, & Hegde, 2013). Consequently, both tagged and shotgun sequencing methods can retrieve extremely rare DNA molecules in a sample, including ones that are in fact environmental contaminants. Furthermore, sequencing is agnostic to the viability of the source material: it does not matter if the DNA came from a living or dead organism (Emerson et al., 2017). Because DNA is stable at room temperature, sterilization techniques kill contaminating microbes but allow their genetic material to persist. Microbial DNA has been regularly recovered from laboratory

reagents, for instance (Kim et al., 2017; Salter et al., 2014) and on sterilized hospital instruments. The combination of ambient DNA and the sensitivity of sequenced-based techniques means that it is extremely common to recover contaminant DNA in metagenomics.

In studies where the goal is to find differentially-enriched microbes in disease, or to determine whether a pathogen is present or not, this contaminant DNA can be a significant confounder. A microbial signature may appear enriched in one disease state over another due only to differences in storage or handling of the samples. Similarly, sequences from a pathogen may appear in idiopathic disease samples, but only be leftover DNA clinging to sterilized hospital equipment. To prevent this, the researcher may employ methods of isolating only DNA from viable organisms such as those described in Emerson et al. (2017), but in many situations collection and storage procedures could kill all the microbes in the sample before workup (such as with formalin fixation). In addition, such methods may bias what is recovered through the metagenomic assay. Regardless of how the viability of the microbes is considered, it is critical to collect appropriate reagent and environmental controls to at least understand the contamination profile of the experiment.

A number of recent studies have described microbial signatures (or even entire microbial communities) in parts of the human anatomy generally considered to be sterile in healthy individuals, including the brain, placenta and in semen (Aagaard et al., 2014; Branton et al., 2013; Hou et al., 2013). Without dismissing these results, it is highly possible that

these microbes came from any number of contaminating sources, including equipment, reagents, or nearby high microbial-load areas such as the mouth or genital tract. All too often, metagenomic studies do not include or show contamination controls, making it hard to assess the validity of their results. In a follow-up study to confirm the presence of microbes in placenta, for instance, Lauder et al. (2016) were unable to see any difference in microbial signatures between placenta samples and reagent controls. Significant work needs to be done from an experimental standpoint before metagenomic sequencing can reach clinically-acceptable levels of confidence.

1.4. Motivations for this thesis

In this thesis, I describe a series of studies that apply metagenomic sequencing to idiopathic and immunological diseases. My aim is to both provide insights into the target diseases and to demonstrate novel experimental methods that increase sequencing precision and sensitivity.

New tools for the identification of etiologic agents

In Chapter 2, I present our efforts to find an etiologic trigger for sarcoidosis, an extensive metagenomic study with broad clinical applications. This is the most comprehensive look at microbial signatures in sarcoidosis to date: we used 16S, ITS, and virome sequencing, as well as shotgun sequencing, in 732 distinct samples over three body sites. We uncovered signatures of a fungus in the Cladosporiaceae family that were enriched in sarcoidosis lymph nodes across two sample cohorts, suggesting a potential etiologic trigger. In addition, this study used a novel experimental design and statistical model that

allowed us to test for enrichment of a microbe in sarcoidosis both in relation to healthy controls and environmental background simultaneously. This experimental design is applicable to other clinical sequencing efforts and represents an advancement in low-biomass metagenomic studies.

Increasing sensitivity of pathogen genomics with *swga*

In Chapter 3, I describe a new program called *swga* for designing primers for use in selective whole-genome amplification (SWGA). SWGA enables the recovery of a target microbe's genome from a complex sample, such as a pathogen or parasite from a host-derived sample. However, the effectiveness of the method relies on the selection of multiple primers that bind preferentially to the target's genome over the background. Identifying an effective primer set is a computationally challenging task that originally involved a lot of manual trial-and-error. The program I describe in this chapter uses an approach derived from graph theory to establish compatible sets of primers for SWGA and evaluates their binding characteristics in arbitrary host/target genomes. I demonstrate how the program and method drastically reduce the sequencing costs for genome recovery of a variety of targets, including *Mycobacterium tuberculosis* and *Plasmodium falciparum* in humans and *Wolbachia pipientis* in *Drosophila*. This program represents an advancement in infectious disease research by enabling researchers to recover the genomes of pathogens that are rare or uncultivable in primary samples. These recovered genomes can then be used to perform population genetics and understand disease outbreaks, or to model genetic variations that affect pathogenicity.

Characterizing immune system and microbiome dynamics in SCID

Finally, in Chapter 4, I integrate metagenomic sequencing to characterize the development of the microbiome in children with severe combined immunodeficiency (SCID) after gene therapy. These children are born without an immune system, but are still colonized by microbes, free of interference from the immune system. After gene therapy, however, the immune system "comes online" and through longitudinal metagenomic sequencing, we can observe how the new immune system and microbiome interact. To characterize the developing immune system, we used sequencing of the CD3 region of the TCR-beta locus in circulating T cells. This allowed us to see clonal outgrowths that suggested the immune system was responding to extant microbiota and correlate T cell dynamics with microbiome changes. Finally, sequencing of gene therapy vector integration sites in the same samples allowed us to estimate the minimum number of cell divisions required to progress from a lymphopoietic progenitor cell to a circulating T cell, increasing our understanding of human immune development.

Taken together, these studies improve our understanding of microbes in disease, provide novel methods for identifying and characterizing pathogens, and help understand the interactions between the immune system and the developing microbiome.

Chapter 2. Microbial lineages in sarcoidosis: A metagenomic analysis tailored for low microbial content samples

The contents of this chapter have been previously published as:

> Clarke, E. L., Lauder, A. P., Hofstaedter, C. E., Hwang, Y., Fitzgerald, A. S., Imai, I., Biernat, W., Rekawiecki, B., Majewska, H., Dubaniewicz, A., Litzky, L. A., Feldman, M. D., Bittinger, K., Rossman, M. D., Patterson, K. C., Bushman, F. D., & Collman, R. G. (2017). Microbial Lineages in Sarcoidosis: A Metagenomic Analysis Tailored for Low Microbial Content Samples. *Am J Respir Crit Care Med*. doi:10.1164/rccm.201705-0891OC

2.1. Abstract

Rationale: The etiology of sarcoidosis is unknown, but microbial agents are suspected as triggers.

Objective: We sought to identify bacterial, fungal or viral lineages in specimens from sarcoidosis patients enriched relative to controls using metagenomic DNA sequencing. Since DNA from environmental contamination contributes disproportionately to samples with low authentic microbial content, we developed improved methods for filtering environmental contamination.

Methods: We analyzed specimens from sarcoidosis subjects (n=93), non-sarcoidosis control subjects (n=72) and various environmental controls (n=150). Sarcoidosis specimens consisted of two independent sets of formalin-fixed, paraffin-embedded lymph node biopsies, bronchoalveolar lavage (BAL), Kveim reagent, and fresh granulomatous spleen from a sarcoidosis patient. All specimens were analyzed by bacterial 16S and fungal ITS rRNA gene sequencing. In addition, BAL was analyzed by shotgun

sequencing of fractions enriched for viral particles, and Kveim and spleen were subjected to whole-genome shotgun sequencing.

Measurements and Main Results: In one tissue set, fungi in the Cladosporiaceae family were enriched in sarcoidosis compared to non-sarcoidosis tissues; in the other tissue set, we detected enrichment of several bacterial lineages in sarcoidosis, but not Cladosporiaceae. BAL showed limited enrichment of *Aspergillus* fungi. Several microbial lineages were detected in Kveim and spleen, including *Cladosporium*. No microbial lineage was enriched in more than one sample type after correction for multiple comparisons.

Conclusions: Metagenomic sequencing revealed enrichment of microbes in single types of sarcoidosis samples, but limited concordance across sample types. Statistical analysis accounting for environmental contamination was essential to avoiding false positives.

2.2. Introduction

Sarcoidosis is a multisystem disease characterized by an aberrant immune response that results in inflammation and granuloma formation. Sarcoidosis is believed to have an antigenic or inflammatory trigger that initiates the immune reaction in a susceptible host (E. S. Chen & Moller, 2014, 2015; Dubaniewicz, 2013). Several susceptibility genes have been identified (Fingerlin, Hamzeh, & Maier, 2015; Fischer et al., 2014) but the trigger remains obscure. Granulomatous inflammation is commonly seen in responses to microbial agents, as are other features of sarcoidosis immunopathology such as

oligoclonal CD4 T cell expansion and TH1 polarization (E. S. Chen & Moller, 2014). No microbial cause has been definitively established for sarcoidosis, but candidates include species of *Mycobacterium* (E. S. Chen et al., 2008; Drake et al., 2002; Dubaniewicz et al., 2007; Song et al., 2005) , as well as fungi (Suchankova et al., 2015) and *Propionibacterium acnes* (Ishige, Usui, Takemura, & Eishi, 1999; Nishiwaki et al., 2004), a common skin bacteria.

The ability to detect rare or unculturable microbes has improved dramatically using deep DNA sequencing (H. Feng et al., 2008; Greninger et al., 2015; Kelly et al., 2016). Several studies have applied bacterial 16S rRNA gene sequencing to sarcoidosis, with differing results (Drake et al., 2002; Richter et al., 1996; Richter et al., 1999). No prior studies have interrogated fungal lineages with tag sequencing, nor used shotgun metagenomic sequencing for comprehensive studies of total DNA or purified viral particles.

We carried out an intensive metagenomic investigation of multiple sarcoidosis sample sets using 16S rRNA gene sequencing to capture bacteria, ITS sequencing for fungi, and whole-genome shotgun sequencing to characterize all microbes. Samples (Table 2-1) include two independent sets of formalin-fixed, paraffin-embedded (FFPE) granulomatous tissue biopsies from newly-identified sarcoidosis patients and controls (sets A and B), bronchoalveolar lavage (BAL) from newly diagnosed untreated Stage II/III sarcoidosis patients and healthy controls (set C), We also interrogated a sample of the Kveim reagent (set D) (which is made from sarcoidosis-affected spleen and was used

historically for sarcoidosis diagnosis by intradermal injection and monitoring for granuloma formation (Klein et al., 1995; Siltzbach, 1961; Teirstein, 1998)), along with fresh granulomatous spleen from a sarcoidosis patient (set E).

An often-underappreciated feature of sequence-based microbial detection is that at low levels of true signal, sequences can be dominated by microbial DNA from environmental sources introduced during sample collection, storage, DNA extraction or other steps (Lauder et al., 2016; Salter et al., 2014). This particularly confounds analysis of samples in which the authentic content of microbial DNA is low, such as lung bronchoscopies and tissue biopsies (Bittinger et al., 2014; Charlson et al., 2011; Robinson, Smith, Sengupta, Prentice, & Sandin, 2013; Salter et al., 2014). Even with the most careful preparation, however, there is often no way to eliminate environmental sequences completely, so further computational and statistical methods must be used to identify contamination. We thus used extensive environmental sampling and applied novel statistical modeling to minimize false positive calls. By investigating several independent sample sets and tissue types, we were able to interrogate whether sarcoidosis-enriched sequences appeared consistently across sample sets. Some of the results of these studies have been previously reported in the form of an abstract (EL Clarke, 2015).

2.3. Methods

2.3.1. Samples collected

2.3.1.1. Archived tissue samples

Two sets of FFPE sarcoidosis and control tissues were analyzed. Set A (from Gdańsk) were mediastinal lymph nodes showing non-caseating granulomas typical of sarcoidosis, and negative by staining for acid-fast or fungal elements. Controls were mediastinal lymph nodes with normal or nonspecific reactive histology. Set B (from Philadelphia) consisted of mediastinal nodes containing granulomas typical of sarcoidosis and negative by fungal and acid-fast stain. Controls were histologically normal nodes from cancer staging procedures. Stored specimens were retrieved and 10um cuts taken under aseptic conditions. Paraffin block environmental controls were cut concurrently with tissue specimens. For set A these were matched from the same block as tissue, while for set B they were not from the same block.

2.3.1.2. Bronchoalveolar lavage (BAL)

BAL fluid (set C) was obtained from subjects undergoing diagnostic bronchoscopy (from Philadelphia) for suspected new diagnosis of pulmonary sarcoidosis who had chest X-rays consistent with parenchymal (Scadding stage II/III) involvement. Subjects included here had sarcoidosis confirmed by standard criteria and exclusion of alternative diagnoses. BAL was performed using standard clinical protocols. Control BAL was obtained from healthy volunteers who underwent research bronchoscopy (Charlson et al., 2011). Prior to bronchoscopy, an environmental control (bronchoscope prewash) was

obtained as previously described (Charlson et al., 2011). BAL and prewash were placed immediately on ice and stored at -80°C until analysis.

2.3.1.3. Kveim and spleen tissue

An aliquot of Kveim reagent ((Teirstein, 1998); set D) was analyzed that was prepared at Mt. Sinai Hospital (New York) for clinical diagnostic use as described (Chase, 1961) and stored under sterile conditions. Sarcoidosis-involved spleen (set E) was obtained from an individual with longstanding disease (from Philadelphia), previously but not currently treated, who underwent splenectomy for symptomatic splenomegaly. Tissue was freshly dissected from the organ and frozen at -80°C. An aliquot of the saline used for tissue homogenization served as a matched environmental control.

2.3.1.4. Human subjects

Tissue samples were obtained from anonymized tissue archives. Bronchoscopy and spleen donor subjects provided written informed consent under IRB-approved protocols.

2.3.2. Analysis

2.3.2.1. Sequence analysis of 16S and ITS rRNA gene segments

Details of extraction, amplification, Illumina sequencing and taxonomic assignment are in Supplement §2.7.1. The bacterial 16S ribosomal RNA gene was amplified using V1V2 primers; this relatively short amplicon was chosen to maximize amplification efficiency for rare sequences from low microbial biomass samples (Charlson et al., 2011; Charlson et al., 2012). The fungal ribosomal RNA internal transcribed spacer ITS1 region was amplified using ITS1F/ITS2 primers (Bittinger et al., 2014; Charlson et al., 2012; Dollive

et al., 2012). Sequences were organized into Operational Taxonomic Units (OTUs) at 97% identity. Statistical analysis was carried out at the individual OTU level, and at genus and family levels.

2.3.2.2. Virome analysis

Virome analysis was carried out on BAL and matched prewash specimens (Abbas et al., 2016; Young et al., 2015). To enrich for viruses, fluid was pelleted and acellular material subject to size-exclusion concentration, followed by nuclease treatment to digest non-encapsulated nucleic acids. Nucleic acids were then extracted, and DNA subjected to whole genome amplification using GenomiPhi. RNA was reverse transcribed to cDNA and PCR-amplified. Resulting libraries were shotgun-sequenced, reads quality filtered, then annotated using a custom database we constructed that included all complete bacterial, fungal, archaeal and viral genomes in RefSeq release 79 (O'Leary et al., 2016). All non-viral reads were removed from consideration. We found many reads annotated to viruses later determined to be either from reagents or mis-annotation of human reads (Abbas et al., 2016), which were therefore excluded. Details are in Supplement §2.7.1.5.

2.3.2.3. Whole genome sequencing

DNA from sarcoidosis spleen tissue and Kveim reagent was subjected to whole genome sequencing (WGS) on an Illumina HiSeq. Reads were quality-filtered, processed, and classified using Kraken (Wood & Salzberg, 2014) with our custom database (described above), with low-complexity regions masked before querying. Details are in Supplement §2.7.1.6. The analytic pipeline is available at https://github.com/eclarke/sunbeam.

2.3.2.4. Accessing sequence data

Sequence data are available in the NCBI SRA under BioProject ID PRJNA392272.

2.3.2.5. Statistical analysis

Code and a complete description are in Supplement §2.7.1.7. For sample sets A and C, which had paired environmental controls, we used the R package *lme4* (Douglas Bates & Walker, 2015) to build a generalized linear mixed effects model (GLMM) to regress the number of reads of a taxa against the study group (sarcoid/healthy) and sample type (tissue/environmental control) (Supp. Figure 2-1). Environmental levels of the taxa in each sample/control pair were captured as a random effect. Enrichment was determined by the significance and directionality of the coefficient for the study group/sample type interaction term after fitting the model. For sample set B, which did not have matched environmental controls, we used the R package *DESeq2* (Michael Love, 2014) to determine enrichment.

Because one could not predict *a priori* whether a putative sarcoidosis-associated microbial trigger would be a specific family, genus, species or even OTU, lineages were tested at the individual OTU level, then aggregated and tested at the species, genus, and family levels. FDR correction was applied at each taxonomic level, and an FDR p-value cutoff of 0.1 was considered significant. While interrogating the data at each taxonomic level increased the risk of type I (false positive) errors, we considered this justified due to uncertainty over which taxonomic level might be linked to sarcoidosis and the exploratory nature of the study, and mitigated by the multiple independent sample sets.

Conversely, since requiring a lineage to reach FDR-corrected significance in multiple independent sample sets would increase the likelihood of type II errors, we also considered lineages that were significant after FDR correction in one sample set, but only significant before FDR correction in other sample sets.

2.4. Results

2.4.1. Sample sets studied

We studied five sets of sarcoidosis samples and controls (Table 2-1). Two (sets A and B) were archival lymph node tissue from patients undergoing diagnostic biopsy, where the sarcoidosis tissue studied was histologically confirmed to show granulomas. Set A included environmental control paraffin blanks matched to the individual tissue block and analyzed in parallel. BAL (set C) was from patients with untreated pulmonary sarcoidosis and healthy volunteers. Reasoning that BAL would most likely reveal a microbial trigger early in the disease course with parenchymal lung involvement, we studied individuals newly presenting with radiological Scadding stage II/III. Environmental controls matched to each sample were prewashes of the bronchoscope used to collect the BAL. We analyzed an aliquot of the Kveim reagent (set D), which is derived from sarcoidosis-affected human spleen and used diagnostically by intradermal injection and monitoring for granuloma formation. Since this suggests an immunological response to a triggering antigen (Klein et al., 1995), we hypothesized that Kveim reagent may contain DNA traces of an etiological microbe. Finally, we tested fresh sarcoidosis-involved spleen (set E), paired with blank controls processed in parallel to model reagent contamination.

Sample Set	*Sample Type*	*Study Group*	*Samples*	*Site*	*Bacteria*	*Fungi*	*RNA viruses*	*DNA viruses*
A	Tissue	Sarcoid	45	Gdansk	643	1180	N/A	N/A
	Tissue	Control	37	Gdansk	207	236	N/A	N/A
	Paraffin only (paired)	Environmental control	82	Gdansk	465	1081	N/A	N/A
	Blanks	Reagent control	27	Gdansk	74	84	N/A	N/A
B	Tissue	Sarcoid	30	Philadelphia	5548	2136	N/A	N/A
	Tissue	Control	19	Philadelphia	2813	2703	N/A	N/A
	Blanks	Reagent control	5	Philadelphia	285	55	N/A	N/A
C	BAL	Sarcoid	16	Philadelphia	3105	25	1	85
	BAL	Healthy subjects	12	Philadelphia	1604	13	1	40
	Prewash (paired)	Environmental control	24	Philadelphia	823	28	4	99
	Blanks	Reagent control	4	Philadelphia	157	22	0	38
D	Kveim reagent	Sarcoid	1	New York	1725	20	N/A	4
	Water	Environmental control	1	Philadelphia	1035	3	N/A	26
E	Spleen	Sarcoid	1	Philadelphia	1156	19	N/A	3
	Saline wash of instruments	Environmental control	2	Philadelphia	408	2	N/A	31
	Water	Reagent control	1	Philadelphia	1035	3	N/A	26

Table 2-1. *Sample sets studied.*

2.4.2. Set A: Lymph node tissue

Microbial lineages detected in set A by bacterial 16S and fungal ITS rRNA gene sequencing are shown as stacked bar graphs (**Error! Reference source not found.**), with d ominant taxa summarized in Supp. Figure 2-2. Each tissue was paired with a control paraffin shaving from the same sample block. Lymph node and environmental control samples are thus plotted side-by-side. In many cases, samples and paraffin controls appear similar.

To investigate community structures in sarcoidosis and healthy lymph node samples, we calculated the UniFrac distance between each pair of samples and tested for clustering using PERMANOVA (Supp. Figure 2-3). Bacterial communities were not significantly

different between sarcoidosis and non-sarcoidosis tissues (Supp. Figure 2-3A), but fungal communities were different (Supp. Figure 2-3B; $p=0.027$, $R2=0.037$). We then asked whether community differences might be attributed to differential contamination. We performed the same PERMANOVA test on paraffin controls from sarcoidosis and non-sarcoidosis samples. No significant difference was detected in bacterial 16S data (Supp. Figure 2-3C), but we did detect a difference in fungal ITS data (Supp. Figure 2-3D; $p<0.002$, $R2=0.091$). Review of the specimen processing pipeline revealed that most sarcoidosis samples (31/45) were stored in a different building from non-sarcoidosis controls. A PERMANOVA test of the effects of storage site on the paraffin environmental controls revealed a significant effect on fungi ($p<0.00001$) but not on bacteria. The environmental fungi responsible for site-specific differences were mostly of the *Aspergillaceae* family (negative binomial test, FDR p-value=0.019; Supp. Figure 2-2).

To account for environmental admixture statistically, we designed a generalized linear mixed model (GLMM) that incorporated each sample's matched environmental control (see Methods). In short, this approach uses the matched control to model the background levels of each taxa. Then, when testing for differential abundance of that taxa, the background levels are accounted for by a separate term in the regression rather than the study group term. We used this approach to test for differential abundance between sarcoidosis and healthy lymph node at the OTU, species, genus, and family level.

Among fungi, at the family level, Cladosporiaceae (within the Capnodiales order; **Error! Reference source not found.**B) was significantly enriched in sarcoidosis (FDR p-value=0.049). At the OTU level, no individual taxa were significantly enriched after FDR correction, but two *Cladosporium* OTUs were significant before FDR correction ($p<0.05$, FDR $p=1$). The Cladosporiaceae fungal lineage is present both in tissue samples and paraffin blank controls, but is most abundant in sarcoidosis tissue (**Error! Reference source not found.**C). No bacterial lineages were significantly enriched in sarcoidosis after accounting for environmental contamination.

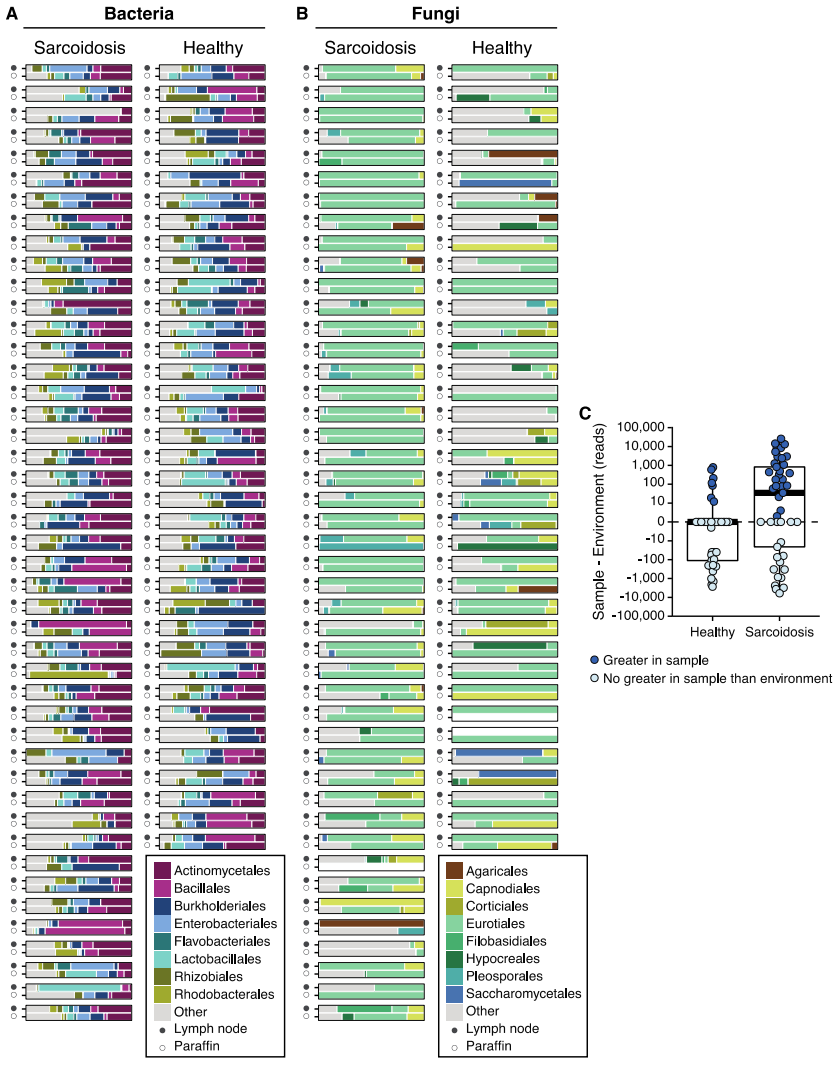


Figure 2-1. Dominant bacterial and fungal orders in lymph node (A).
The major bacterial (A) and fungal (B) orders identified by 16S and ITS rRNA gene sequencing, respectively, are shown as proportions of total reads. Less common lineages are aggregated under "Other." For each pair, the closed symbol (●) indicates the FFPE lymph node sample, while the open symbol (○) indicates a slice of blank paraffin cut from the same block to serve as an environmental control. Empty (white) bar charts indicate that the sample was either not available or had no detectable lineages. The difference in Cladosporiaceae reads between a sample and its environmental control are shown in (C). Closed circles represent samples with more Cladosporiaceae reads in the sample than the matched environmental control, while open circles represent samples in which the number of Cladosporiaceae reads were not greater than in the environment control. The abundances are shown as reads to more accurately reflect the input to the test, which used raw read counts as input. Normalization between differing sequencing depths was accounted for by modeling library size as a random effect for each sample (see Methods).

2.4.3. Set B: Lymph node tissue

The dominant bacterial and fungal lineages in tissue set B are shown in Figure 2-2, with rank abundance plots in Supp. Figure 2-4. We compared community structure using UniFrac and PERMANOVA (Supp. Figure 2-5), and found differences in bacterial (but not fungal) populations between the sarcoidosis samples and healthy controls ($p<0.05$). We then tested for differentially abundant taxa at the OTU, species, genus, and family levels. Numerous bacterial taxa were significantly enriched in sarcoidosis compared to control tissues, including three OTUs in the *Corynebacterium* (order Actinomycetales) genus (FDR p=1e-5, 0.064 and 0.067, respectively) and four OTUs in the Rhodocyclaceae family (order Rhodocyclales, FDR p=0.076, 0.003, 2e-05, and 0.005, respectively). Other sarcoidosis-enriched bacteria were from the Sphinogomonadaceae family (order Sphingomonadales), the Comamonadaceaea and Oxalobacteraceaea families (order Burkholderiales), and the Moraxellaceae and Pseudomonadaceae families (order Pseudomonadales). No fungal lineages were sarcoidosis-enriched in tissue set B after FDR correction, including *Cladosporium* (although fungi of this family do appear to be present in higher levels in the sarcoidosis samples; Supp. Figure 2-4). Given the absence of paired environmental controls, these results are limited in isolation and serve mainly for comparison with other sample sets.

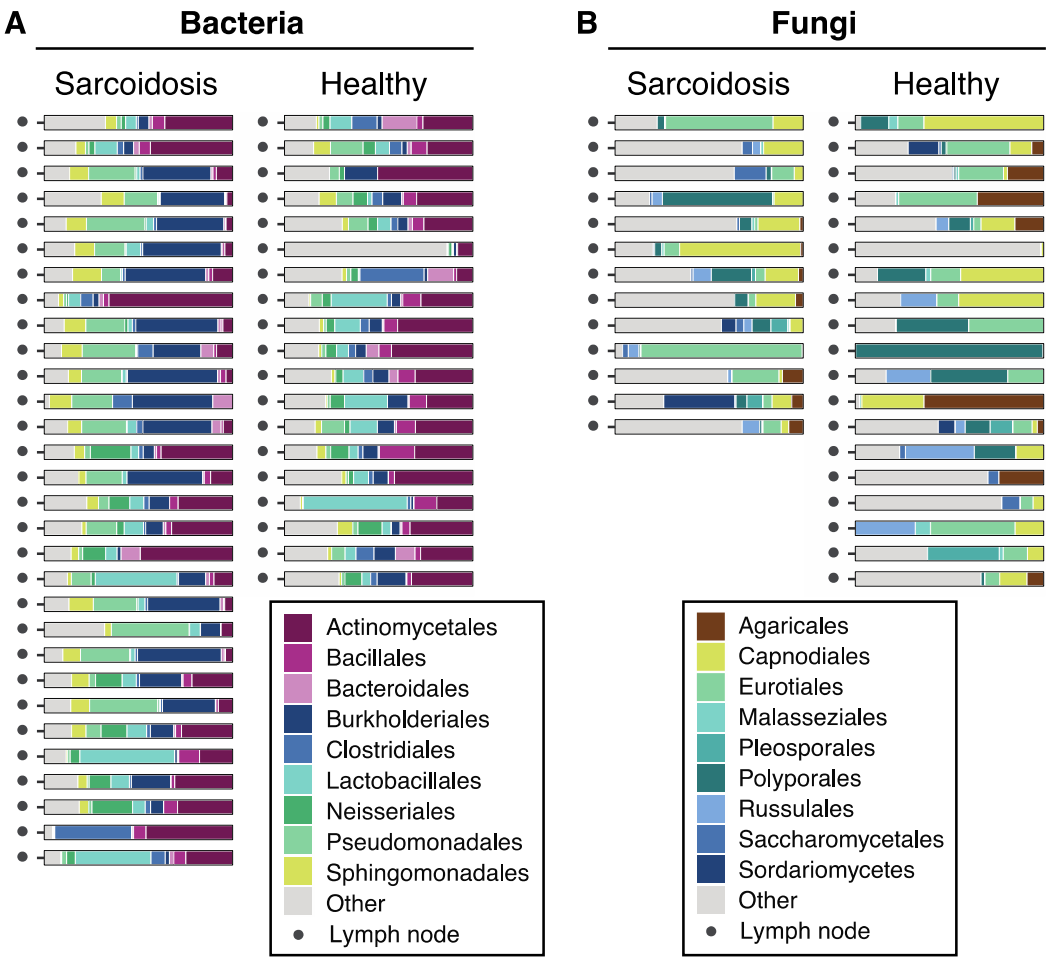


Figure 2-2. Bacterial and fungal lineages in lymph node (B).
The major bacterial (A) and fungal (B) orders identified by 16S rRNA and ITS gene sequencing are shown as proportions of total reads. Less common lineages are aggregated under "Other." Seventeen samples failed to amplify any usable ITS sequences in B and are omitted. Blank paraffin controls matched to each tissue specimen were not available for these samples.

2.4.4. Set C: Bronchoalveolar lavage

We analyzed DNA from whole BAL for bacteria and fungi using 16S and ITS gene sequencing (**Error! Reference source not found.**, Supp. Figure 2-6), along with matched bronchoscope pre-washes. Analysis using UniFrac and PERMANOVA (Supp. Figure 2-7) showed no significant differences between sarcoidosis and control bacterial

communities. To identify taxa enriched in sarcoidosis while accounting for environmental input, we employed the GLMM described above. We found that the bacterial family Corynebacteriaceae (order Actinomycetales) was enriched in sarcoidosis before FDR correction, but no taxa were enriched after FDR correction.

Fungal sequences in BAL were sparse (**Error! Reference source not found.**B), c oncordant with previous reports on BAL fungal detections (Bittinger et al., 2014). However, the genus *Aspergillus* (within the Eurotiales order; **Error! Reference source n ot found.**B) was enriched in sarcoidosis (FDR p=0.042).

2.4.4.1. Virome analysis of sarcoidosis BAL

We investigated the lung virome in sarcoidosis by generating virus particle preparations from acellular BAL and matched prewashes, and deep-sequencing both RNA and DNA. Initial inspection revealed abundant reads annotated as HHV6/HHV7. These reads matched human simple sequence repeats (Abbas et al., 2016), and were therefore removed. We also purged sequences that were present in blank controls and attributable to viral enzymes used as reagents, and thus likely reagent-derived. Our approach was designed to detect both DNA and RNA viruses, but we did not recover any RNA viruses that did not likely originate from reagent contamination.

The majority of remaining viral sequences were phages of the Siphoviridae and Iridoviridae lineages (**Error! Reference source not found.**C). The data initially s uggested a much richer population of viruses in sarcoidosis BAL than healthy controls. However, viral sequences in the sarcoidosis cohort's prewash controls were also richer

than the control cohort's prewash (Supp. Figure 2-8). This difference is likely because sarcoidosis subjects underwent bronchoscopy in a clinical endoscopy suite, whereas healthy volunteers were sampled in a different facility used for research studies. This suggests that each location contributed a different environmental background of virus sequences, likely originating in lavage saline or water used to rinse bronchoscopes after cleaning.

We therefore used the same GLMM approach to account for environmental differences when testing for enriched viral species. No viruses were sarcoidosis-enriched at any taxonomic levels tested. Without accounting for environmental input, the enrichment analysis would have been confounded by differences resulting from bronchoscopy locations.

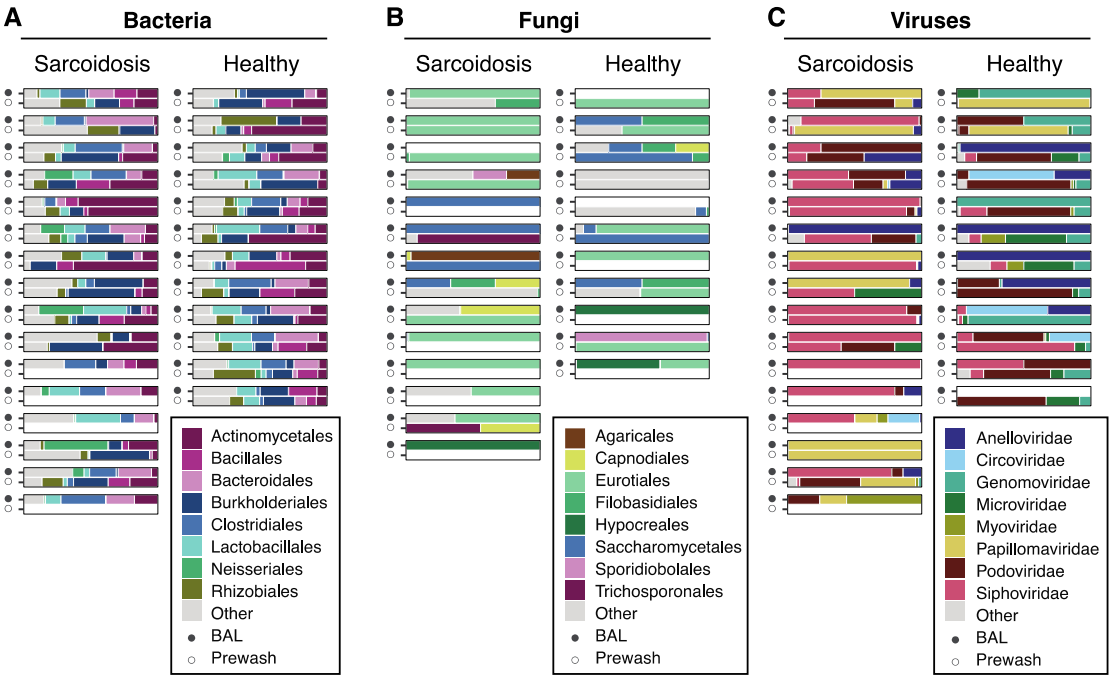


Figure 2-3. Bacterial, fungal and viral lineages in BAL.
The major bacterial (A) and fungal (B) lineages identified by 16S rRNA and ITS gene sequencing, and viral (C) lineages identified by shotgun sequencing of all nucleic acids in virus particle preparations, are shown as proportions of the total reads. Data are shown at the order level for A and B, and the family level for C. Less common lineages are aggregated under "Other." For each pair, the closed symbol (●) indicates the BAL fluid, while the open symbol (○) represents the prewash fluid for that scope. Empty (white) bar charts indicate that the sample was either not collected or had no detectable lineages. Three sample pairs failed to amplify any ITS sequences and are omitted from B.

2.4.5. Sets D and E: Kveim reagent and sarcoidosis spleen

We analyzed Kveim reagent and fresh spleen from a patient with sarcoidosis. Three separate pieces of spleen were tested, along with controls to capture sequences from the environment. Kveim, spleen and controls were subject to 16S and ITS sequence analysis (Figure 2-4A and B) and also shotgun whole-genome sequencing (WGS) (Figure 2-4C). WGS yielded mostly human sequences, which were removed; the remaining sequences queried for microbial annotations.

The predominant bacteria found by both 16S and WGS were in the Propionibacteriaceae family (within the Actinomycetales order), and were detected across all samples including controls. Other ubiquitous taxa included Corynebacteriaceae and Pseudomonadaceae (of the Actinomycetales and Pseudomonadales orders, respectively). Some differences were seen between 16S and WGS analysis for other taxa, likely resulting from the relative representation of sequences within 16S and WGS databases. No taxa were present only in sarcoidosis samples and not environmental controls.

Fungal detections were sparse in both ITS sequencing and WGS, and inconsistent between methods. Cladosporiaceae (order Capnodiales) was detected by ITS in one spleen sample, but not by WGS. This may be due to a paucity of database genomic sequences for Cladosporiaceae, limiting detection in WGS annotation. There was no consistent fungal detection in sarcoidosis samples versus controls.

In the WGS data, we initially detected alignments annotated as *Toxoplasma gondii* in the Kveim and spleen samples. We also detected reads annotating to an unfinished *Mycobacteria* genome. However, these sequences were found to match human microsatellite simple repeats, and so were judged to be false-positives and removed (detailed in Supplement §2.7.1.5). This is an issue for WGS data but not for 16S or ITS analysis, as the untargeted approach allowed capture of low-complexity repeat DNA. The only viral reads detected were from bacteriophages and were also found in the controls, and thus inferred to be environmentally-derived.

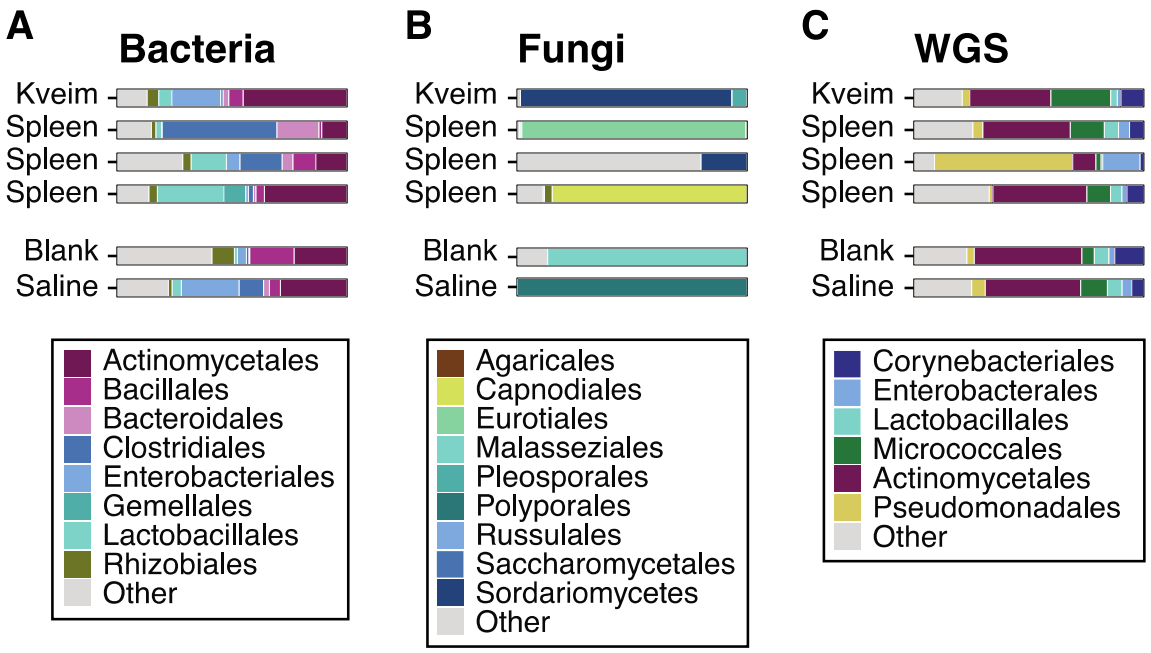


Figure 2-4. Microbial lineages in Kveim and sarcoid spleen.
The major lineages in sample sets D (Kveim) and E (sarcoidosis spleen) shown by sequencing. (A) Shows results from 16S sequencing, (B) shows ITS sequencing, and (C) shows results from whole-genome shotgun sequencing, after filtering as described in Supplemental Methods §2.7.1.6. Less common lineages are aggregated under "Other", including fungal detections in (C).

2.4.6. Shared lineages

No bacterial or fungal lineages were significantly enriched after FDR correction in more than one sample set. To broaden our search, we examined lineages that were significantly enriched in one sample set after FDR correction, and queried their abundance in the other sets. Enriched lineages are summarized in Table 2-2.

Sample Set	*Kingdom*	*FDR p < 0.1*	*non-FDR p < 0.05*
Tissue Set A	Bacteria	None	None
	Fungi	Cladosporiaceae	OTU6408 (genus Cladosporium), Cladosporiaceae
Tissue Set B	Bacteria	Many (113), including Corynebacterium and Rhodocyclaceae	Many (252)
	Fungi	None	Many (149), including one Cladosporium OTU (OTU7142)
BAL Set C	Bacteria	None	OTU 104987 (Family Rhodocyclaceae), OTU 4301737 (genus Porphyromonas), Corynebacterium, Neisseria
	Fungi	Aspergillus	Aspergillus
	Viruses	None	None

Table 2-2. Summary of taxa enriched in sarcoidosis.
Microbial taxa enriched in sarcoidosis lymph node and BAL over healthy controls.

In tissue set A, fungi from the Cladosporiaceae family (order Capnodiales) were significantly enriched in sarcoidosis when tested as a group. A single *Cladosporium* OTU (OTU7142) was enriched in tissue set B before multiple testing correction (p=0.042), though not the Cladosporiaceae family overall. Cladosporiaceae were detected but not enriched in sarcoidosis BAL. Finally, abundant *Cladosporium* reads were detected in one of three replicate spleen samples via ITS sequencing and not in the environmental controls, although it was not in WGS of spleen or Kveim.

In tissue set B, three OTUs belonging to the *Corynebacterium* bacterial genus (order Actinomycetales) were significantly enriched in sarcoidosis. While no individual *Corynebacterium* OTUs were enriched in other sample sets, the *Corynebacterium* genus was enriched in sarcoidosis BAL before FDR correction (p=0.02). *Corynebacterium* were detected but not sarcoidosis-enriched in tissue set A, and also detected in Kveim and spleen, as well as environmental controls.

Also in tissue set B, multiple OTUs in the Rhodocyclaceae (order Rhodocyclales) bacterial family were significantly enriched. A single Rhodocyclaceae OTU was enriched in BAL before FDR correction (OTU104987, genus *Hydrogenophilus*, p=0.009). No Rhodocyclaceae lineages were detected in tissue set A, but appeared in both Kveim and spleen as well as controls from sets D and E.

In BAL (set C), fungi in the *Aspergillus* genus (order Eurotiales) were significantly enriched in sarcoidosis. *Aspergillus* was detected in tissue set A, but was not sarcoidosis-enriched. Numerous *Aspergillus* lineages were also detected but not sarcoidosis-enriched in tissue set B. *Aspergillus* species were not detected in Kveim or spleen in sets D and E, but were found in the environmental and blank controls by WGS.

2.5. Discussion

This is the first study to interrogate microbial agents in sarcoidosis using a metagenomic approach combining bacterial and fungal sequence tag analysis, virome shotgun sequencing, and whole genome sequencing. We anticipated that a causal microbe would be present in low abundance, so rigorous consideration of potential contamination would be critical for distinguishing authentic from environmentally-derived sequences. Our findings were inconsistent across the five sample sets analyzed (Table 2-1, Table 2-2), but do provide candidates for further validation, and strongly emphasize the importance of assessing environmental contamination.

Cladosporiaceae was significantly enriched in sarcoidosis specimens in tissue set A after adjustment for environmental admixture and multiple comparisons, and also appeared in several other sample sets, though not with comparable statistical enrichment. Fungi in the Cladosporiaceae family are extremely common in the environment (Ezike, Nnamani, Ogundipe, & Adekanmbi, 2016; Peternel, Culig, & Hrga, 2004), can trigger hypersensitivity pneumonitis and asthma (Chiba et al., 2009; Tham et al., 2017), and are capable of eliciting granulomatous inflammation (Robinson et al., 2013; Silva & Ekizlerian, 1985). This finding may warrant further investigation.

In tissue set B, we detected multiple sarcoidosis-enriched taxa, but interpretation is limited by the lack of matched environmental controls. Enriched taxa included several *Corynebacterium* OTUs, which were also sarcoidosis-enriched before FDR correction in BAL (set C). Similarly, OTUs annotated as Rhodocyclaceae were significantly enriched in set B, and enriched before FDR correction in set C. *Corynebacterium* are particularly interesting because they are known to elicit granulomatous responses in vivo (Nureki et al., 2007; Taylor, Paviour, Musaad, Jones, & Holland, 2003), although the association with sarcoidosis in this study was weak.

In addition to histopathological similarities, mycobacteria have been linked to sarcoidosis by immunological responses and/or sequence-based detection (E. S. Chen et al., 2008; Drake et al., 2002; Dubaniewicz et al., 2007; Song et al., 2005). However, we did not find enrichment of mycobacteria in sarcoidosis. Mycobacteria are difficult bacteria to isolate DNA from due to tough cell walls. To ensure our methods were robust, we confirmed

detection via sequencing in known mycobacteria-infected tissue samples, and biological specimens spiked with avirulent *M. tuberculosis* (not shown). We also found low levels of mycobacteria in many samples and environmental controls. This suggests that our methods are not inherently insensitive to mycobacteria, but that mycobacteria as a group were not enriched in these sarcoidosis specimens. However, the 16S variable region amplified, V1V2**,** cannot distinguish between mycobacterial species, which precludes detection of species-level differences. We also found that *Propionibacterium acnes* was a ubiquitous environmental agent, concordant with other studies (Mollerup et al., 2016), with no evidence of enrichment in sarcoidosis.

Environmental admixture is an issue in any metagenomic survey and becomes increasingly important as the amount of authentic microbial content decreases (Salter et al., 2014). Such sequences can be introduced from specimen collection and storage, DNA extraction kits, the processing pipeline, or even "barcode error" inherent in Illumina deep sequencing platforms that can allow low-level bleed-over in the sequencing process (Lauder et al., 2016). Most importantly in studies comparing subject groups, clinical samples collected at different times or locations may be contaminated with different environmental sequences. This is especially problematic when taxa of interest may also be environmentally present, so simple subtraction of background lineages is inappropriate. For example, sarcoidosis and healthy BALs were acquired in different locations and showed different background viromes. For tissue set A, specimen storage location differed between study groups, which led to enrichment of environmental fungi in one group and not the other.

A key component of our approach is the generalized linear mixed model, which enabled us to capture and control the effects of differential environmental admixture without losing the ability to test for differential abundance in environmental taxa. In tissue set A, without accounting for environmental input, a naïve enrichment analysis would have identified fungal species within the *Aspergillus* and *Penicillium* genera as sarcoidosis-enriched. The same is true for viruses in BAL (set C), which would have incorrectly identified greater phage populations in sarcoidosis. The GLMM approach presented here would enable handling of potential confounding effects of environmental admixture in microbiome studies generally, and is particularly critical for specimens with low authentic microbial content, when coupled with appropriate matched environmental controls for each clinical sample.

Our study has several limitations. We investigated microbes that might be enriched in sarcoidosis at time of diagnosis, which is the earliest time point feasible, but the time from actual disease onset is unknown, so an etiological trigger may no longer be present. Conversely, it is conceivable that microbial enrichment associated with sarcoidosis could be a consequence of the disease, rather than a cause. Use of samples from distinct geographic locations would reveal shared lineages, but sarcoidosis triggers may differ geographically. Additionally, any triggers may not be enriched in sarcoidosis subjects at all, but may be ubiquitously present, with disease determined mainly by host susceptibility factors. Although we examined a total of 93 sarcoidosis and 72 non-sarcoidosis specimens, plus 150 environmental controls (for a total of 738 sequencing reactions) the number of samples in any one set was modest. For the FFPE samples,

sensitivity may be lessened by damage DNA incurred by during the de-crosslinking step necessary to undo the formalin fixation (Campos, 2011). For our DNA virus methods, Genomiphi amplification may introduce bias towards short circular DNA due to rolling-circle amplification, although this bias should be consistent across study groups. Finally, any primers chosen for tagged sequencing also have inherent biases and may be more sensitive to some microbes than others (such as with the V1V2 primers and *Mycobacterium* species, as discussed previously).

In summary, application of metagenomic sequencing and analytic approaches tailored to low microbial-biomass samples did not identify a single causative agent but identified several candidate agents as sarcoidosis-enriched. These include the Cladosporiaceae fungal family and *Corynebacterium* bacterial taxa. The modest enrichment and limited concordance of these candidates in the sample sets precludes our ability to assert any causal relationship with sarcoidosis, but we believe these candidates may be of interest in future studies. More broadly, the model we present here increases the power of metagenomic studies in low microbial biomass samples, such as lung and tissue specimens, by allowing researchers to account for and test environmental admixture, thus avoiding potential spurious identifications.

2.6. Acknowledgements

We are grateful to subjects who generously volunteered for this study, to M. Padilla at Mount Sinai for the donation of Kveim reagent, to A. Ziober for assistance with tissue specimens, to W. Drake for avirulent mycobacteria, and to members of the Collman and

Bushman laboratories for help and suggestions. This work was supported by the NIH grants U01HL112712 (Site-Specific Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis (GRADS) Study) and R01HL113252, and received assistance from the Penn Center for AIDS Research (P30AI045008) and the Penn-CHOP Microbiome Program.

2.7. Supplemental Material

2.7.1. Supplemental Methods

2.7.1.1. Specimens analyzed

Formalin-fixed paraffin-embedded (FFPE) sarcoidosis and control tissues in set A were from the Medical University of Gdansk (Gdansk, Poland) and tissue set B were from the Hospital of the University of Pennsylvania (Philadelphia, PA, USA). Paraffin block environmental controls targeted a region of the block that did not contain tissue, and were cut at the same time as the tissue specimens.

Bronchoalveolar lavage (BAL) fluid (set C) was obtained from subjects with Scadding stage II or III chest X-rays undergoing diagnostic bronchoscopy for a suspected new diagnosis of pulmonary sarcoidosis. Subjects included here are those in whom sarcoidosis was confirmed based on standard criteria including transbronchial biopsy with noncaseating granulomas and exclusion of alternative diagnoses by culture and stains for fungi and mycobacteria. All subjects were newly-diagnosed and not previously treated. Non-sarcoidosis control BAL was obtained from healthy volunteers who underwent research bronchoscopy and have been described previously (Charlson et al.,

2011). An environmental control (bronchoscope prewash) was obtained prior to each bronchoscopy by suctioning 10 ml of lavage saline through the scope channel as previously described (Charlson et al., 2011).

An aliquot of Kveim reagent (set D) was analyzed that was previously prepared at Mt. Sinai Hospital and validated for clinical diagnostic use as described (Siltzbach, 1961; Teirstein, 1998).

A specimen of sarcoidosis-involved spleen (set E) was obtained from an individual with sarcoidosis who underwent splenectomy for symptomatic splenic enlargement. Histological examination subsequently confirmed granulomatous involvement of the spleen. Immediately following surgical removal, subcapsular tissue was dissected from the organ under aseptic conditions, transported in a sterile container on ice, cut into 0.1g pieces with sterile scissors and forceps under sterile conditions, then snap-frozen and stored at -80ºC. For analysis, one 0.1g piece of tissue was thawed, cut in thirds with a sterile scalpel, and each fragment subject to independent DNA extraction and analysis in parallel. To obtain an appropriate environmental control that reflected both specimen processing and sequencing steps, prior to tissue dissection an aliquot of saline (that later served as a vehicle for tissue processing) was used to gently rinse the scalpel that was subsequently employed for tissue dissection, and this was carried through the DNA extraction and sequencing pipeline.

2.7.1.2. DNA purification

DNA from paraffin-embedded, formaldehyde-fixed tissue samples (10um slices) was extracted using the Qiagen GeneRead DNA FFPE Tissue kit following manufacturer's recommendations, except with the addition of a 10 minute, 95ºC incubation step during proteinase digestion to maximize DNA yield from hard-to-lyse fungi and other microbes, and increase DNA yield. For 16S and ITS sequencing, DNA from 1.8 ml of unfractionated BAL fluid and the corresponding scope prewashes was isolated using the PowerSoil DNA kit (MoBio, Carlsbad, CA), and included an additional 10 min, 95ºC incubation step as above. DNA and RNA isolation for virome sequencing is described below. For spleen, three sections of approximately 0.03g each were dissected from a larger 0.1g piece under sterile conditions, and DNA was isolated using the Qiagen QIAamp Pathogen UCP Mini kit. For Kveim reagent, 200ul aliquots of material were spun down at 10,000 RPM in a tabletop centrifuge for 10 minutes and the supernatant was discarded. The pellet was resuspended in SM buffer and extracted using the Qiagen QIAamp Pathogen UCP Mini kit. All extractions were performed in a BSL2+ hood after the workspace was treated with DNA remover and UV irradiation to remove environmental contamination. DNA was stored at -20ºC.

2.7.1.3. Sequence analysis of 16S and ITS rRNA gene segments

Bacterial 16S ribosomal DNA was amplified using primers for the V1V2 16S region (Supp. Table 2-1) and PCR conditions as described previously (Charlson et al., 2012; Lauder et al., 2016). Each PCR reaction was carried out in duplicate or triplicate in 25ul reactions and pooled before sequencing. The PCR reactions were conducted using

AccuPrime Taq DNA Polymerase from Invitrogen and the pooled amplicons were sequenced on an Illumina MiSeq. Resulting sequence reads were processed using the QIIME 1.91 workflow (Caporaso, Kuczynski, et al., 2010). In brief, the reads were clustered into OTUs with 97% sequence similarity using UCLUST (Edgar, 2010) and aligned to full-length 16S sequences using pyNAST (Caporaso, Bittinger, et al., 2010). Taxonomic ranks were assigned using RDP Classifier (G. P. Wang, Ciuffi, Leipzig, Berry, & Bushman, 2007) with minimum 50% confidence.

The fungal ITS1 region was amplified using ITS1F and ITS2 primers ((Dollive et al., 2012; Gardes & Bruns, 1993), Supp. Table 2-1), with each sample individually barcoded using Golay barcodes. The PCR reactions were carried out in duplicate or triplicate with 4ul of template, 0.4 ul AccuPrime Polymerase, 3ul of 10uM forward primer, 3ul of 10uM reverse primer, 2.5 ul Buffer II, and 12.1 ul PCR-grade water. The reactions were conducted using cycling parameters as follows: 94^{o}C initial denaturation for 3 minutes; 94^{o}C for 45s, 56^{o}C for 60s, 72^{o}C for 90s (35 cycles); 72^{o}C final extension for 10 minutes. The individual replicates were bead purified using Agencourt AMPure XP beads (1:1 ratio), and then purified a second time with a 0.8 ratio to remove excess primer dimers. Amplicon concentration was assessed using Picogreen and product size checked on a BioAnalyzer. The final products were pooled for sequencing and bead-purified again at a 0.8 ratio to remove further primer dimers. The sequencing was performed on an Illumina Miseq. After sequencing, the reads were processed using PIPITS (Gweon et al., 2015). Taxonomy was assigned using BROCC (Dollive et al., 2012) and all subsequent analysis

was performed in R. All synthetic DNA sequences used in this study are in Supp. Table 2-1.

2.7.1.4. Virome analysis

To enrich for viruses in BAL and matched prewash specimens, the following steps were used: 1.8ml of BAL fluid was pelleted at 960g for 10 min and the acellular supernatant material then subjected to size exclusion concentration (100 kDa; Amicon). The filtered material was then nuclease treated to digest non-encapsulated nucleic acids. Nucleic acids were extracted, DNA subjected to whole genome amplification using GenomiPhi, and RNA was transcribed to cDNA and PCR-amplified. Details of these methods have been previously described (Abbas et al., 2016). The resulting libraries were shotgun sequenced on an Illumina HiSeq 2500 using the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) with dual-indexed barcodes, and reads were quality filtered using Trimmomatic (Bolger, Lohse, & Usadel, 2014).

Human reads were filtered by removing any that mapped to the human genome (GRCh38). Remaining reads were annotated using Kraken (Wood & Salzberg, 2014) using a custom database that included all complete bacterial, fungal, archaeal and viral genomes available in RefSeq release 79 (O'Leary et al., 2016). All non-virus reads were removed from consideration. We found many reads annotated to viruses later determined to be either from reagents or otherwise spurious, including mis-annotation of human reads, as we have previously reported (Abbas et al., 2016). We excluded the following species: Enterobacteria phage M13, Enterobacteria phage T7, Enterobacteria phage phiX-

174 sensu lato, Bacillus phage phi29, and Pseudomonas phage phi6, human herpesvirus 6 and 7, and Shamonda virus.

2.7.1.5. Additional quality control issues

To minimize the impact of batch effects (Salter et al., 2014), a single lot of DNA extraction kits were used for all samples of a given type (including sarcoidosis, non-sarcoidosis, and evironmental controls), and all samples of each set were combined and sequenced in a single sequencing run. Because fixation can partially degrade DNA and subsequent downstream analysis (Campos, 2011), we used the Qiagen GeneRead FFPE kit, which includes an enzyme to correct C->T mutations that may occur. This kit does not address increased DNA fragmentation from fixation, but we expect these effects to be minimal due to the small length of the target V1V2 and ITS1 regions in 16S and ITS sequencing.

2.7.1.6. Whole genome sequencing

Sarcoid spleen tissue and Kveim reagent were analyzed using metagenomic whole genome shotgun sequencing. DNA was extracted using the Qiagen Ultraclean Pathogen (UCP) Mini kit following its recommendations for DNA isolation from tissue samples. Metagenomic sequencing was carried out on an Illumina HiSeq 2500 using the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) with dual-index barcodes. The reads from the metagenomic sequencing were then processed in the following steps: 1) reads were quality-filtered, paired and adapter-trimmed using Trimmomatic (Bolger et

al., 2014); 2) human and phiX reads (used in sequencing library prep) were removed using bwa (Li & Durbin, 2009), and 3) reads were classified using Kraken (Wood & Salzberg, 2014) with a custom database built from all genomic sequences from RefSeq (release 79, (O'Leary et al., 2016)), with low-complexity regions masked before querying. The complete pipeline is available at https://github.com/eclarke/sunbeam.

In initial analysis, reads containing short sequence repeats from human DNA were annotated to various species including *Toxoplasma gondii* and *Mycobacterium* spp. The short repeat sequences in these reads were also present in some draft genomes used to build our database, and were sufficiently complex to avoid masking by the NCBI *dust* program (Camacho et al., 2009) used on all database sequences. These annotations were judged to be false positives based on their presence in the human genome and the lack of any other reads aligning to *Toxoplasma gondii* and *Mycobacterium* spp.. To prevent further false positives, we used the *RepeatMasker* program (Smit, 2013-2015) to mask these repeat regions and re-ran the classification. The results presented here reflect classification after repeat masking.

2.7.1.7. Statistical analysis

In order to compare study group samples while accounting for the environmental controls, we developed a generalized linear mixed effects model (GLMM) with the following design. The presence or absence of a taxa was modeled using a binomial link function, and the fixed effects were study group (sarcoidosis or healthy) and sample type (BAL vs prewash, or FFPE vs blank paraffin). The random effects were the grouping pair

(i.e. the sample with its matched control) plus random effect for each individual sample. The inclusion of this latter random effect helped control overdispersion (Harrison, 2014) and account for varying library size between samples. The model was built using the 'glmer' function in the R package *lme4 (Douglas Bates & Walker, 2015)*. To determine if a taxon was significantly enriched, we looked at the significance and directionality of the coefficient on the study group and sample type interaction term. Specifically, the coefficient had to be positive in the sarcoidosis and non-environmental direction to be considered enriched in sarcoidosis over both the healthy and environmental background. The significance of the interaction coefficient was measured both by testing the model with and without the interaction term via ANOVA. The significance of the interaction coefficient was measured by ANOVA, comparing to an alternative model without the interaction term. Taxa appearing in less than 10% of the samples, for which the model failed to fit, or which had a negative interaction coefficient after fitting were discarded from further consideration. P-values from the remaining taxa were subjected to multiple testing correction using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). We set our FDR-corrected p-value threshold at 0.10 to prioritize finding potential hits.

For tissue cohort B, we did not have matched environmental controls, and so the GLMM specified above was unnecessary. Instead, we used the R package *DESeq2* (Michael Love, 2014) to assess enrichment of a taxa in sarcoidosis samples over healthy controls. DESeq2 uses a negative binomial distribution to fit taxon abundance in samples in a generalized linear model, but cannot model random mixed effects. The package provides p-value and multiple testing correction automatically, and for consistency we used the

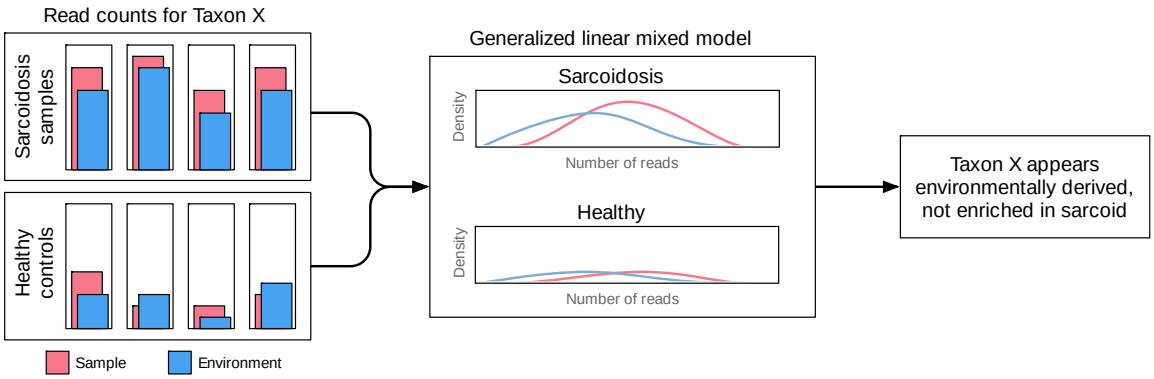
same FDR threshold of 0.1. DESeq2 is one of the currently recommended methods for testing for differentially abundant taxa by the QIIME developers (Caporaso, Kuczynski, et al., 2010).

In both the GLMM or the DESeq2 approach, we tested at a hierarchy of taxonomic ranks. After testing each individual OTU, we collapsed the counts according to species so that the reads of all OTUs belonging to the same species were summed together. We then tested each species from the same model, and repeated this process for genus and family level ranks. Our rationale for this was that etiologic agents may be multiple species or taxa within a higher group- e.g. a family of molds, or a number of closely related bacterial species.

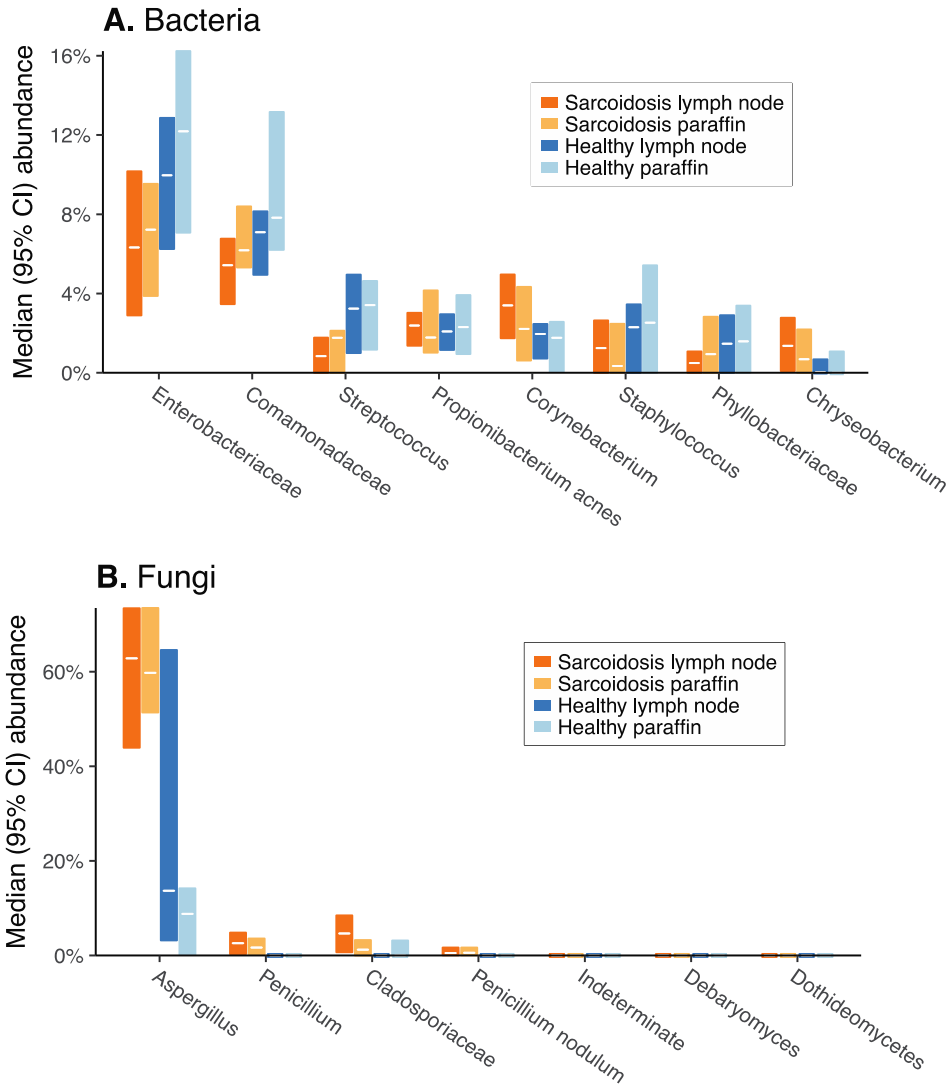
2.7.1.8. Data availability

The code used for all analysis, including the specific model formulation, is available online at https://github.com/eclarke/sarcoid-microbiome-paper. Post-processed sequence data (after quality control, OTU formation, and taxonomic assignment) is archived at https://zenodo.org/record/825276. Raw sequence reads are archived with NCBI under BioProject PRJNA392272.

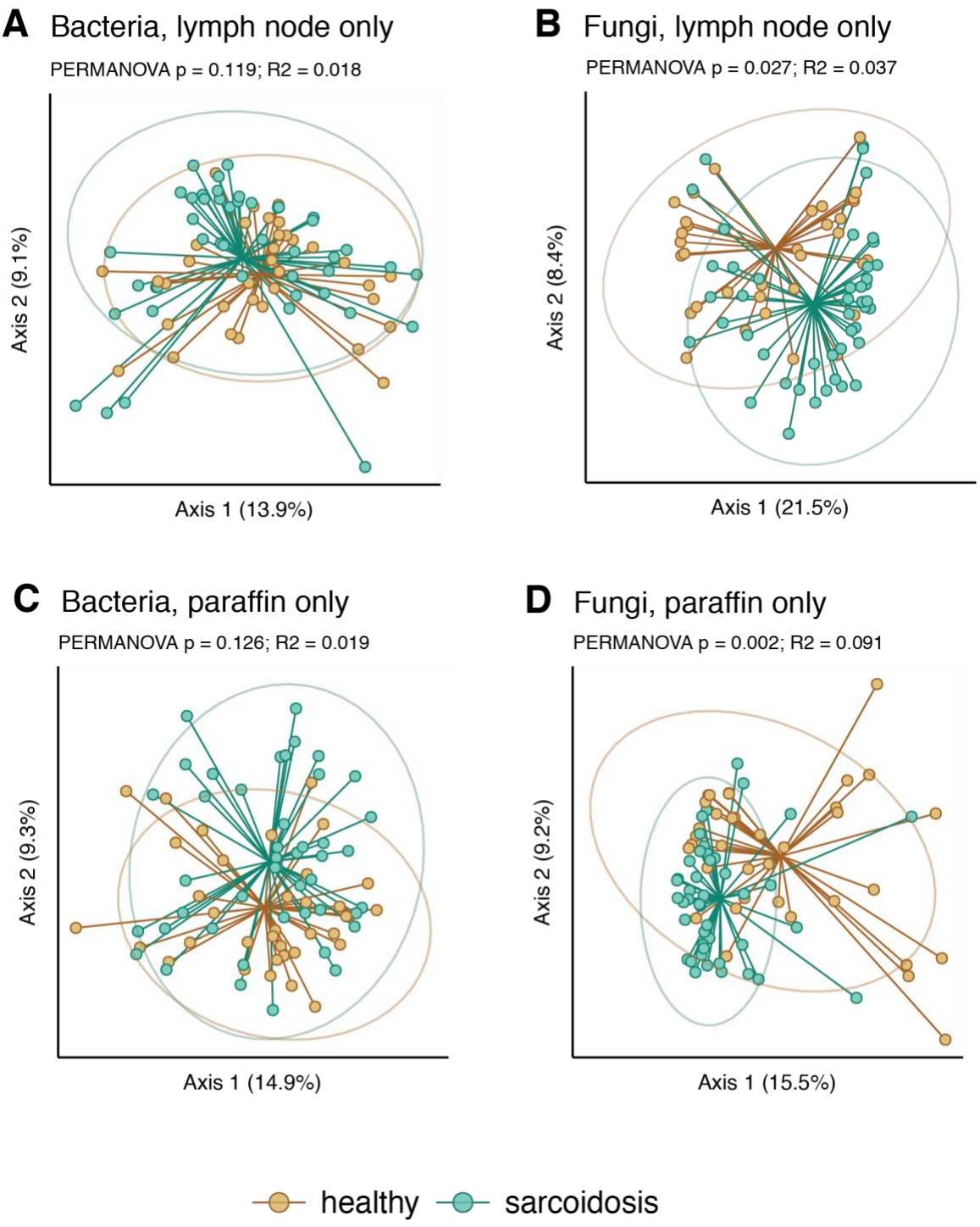
2.7.2. Supplemental Figures



Supp. Figure 2-1. Illustration of statistical approach.
Each taxa is considered individually. Background reads of the taxa from its environmental control (blue) are considered with the reads detected in the paired sample (red). The reads from a taxon are modeled using a binomial distribution link function in a generalized linear mixed model, and enrichment is determined by the magnitude and direction of the regression term corresponding to sarcoidosis samples when contrasted with healthy samples and environmental controls. A positive coefficient for the sarcoidosis + sample type interaction term indicates enrichment of that taxa relative to environment and healthy controls; statistical significance is assessed via ANOVA with a model lacking the interaction term between sample type and study group. The code describing the model exactly is available online at https://github.com/eclarke/sarcoid-microbiome-paper.

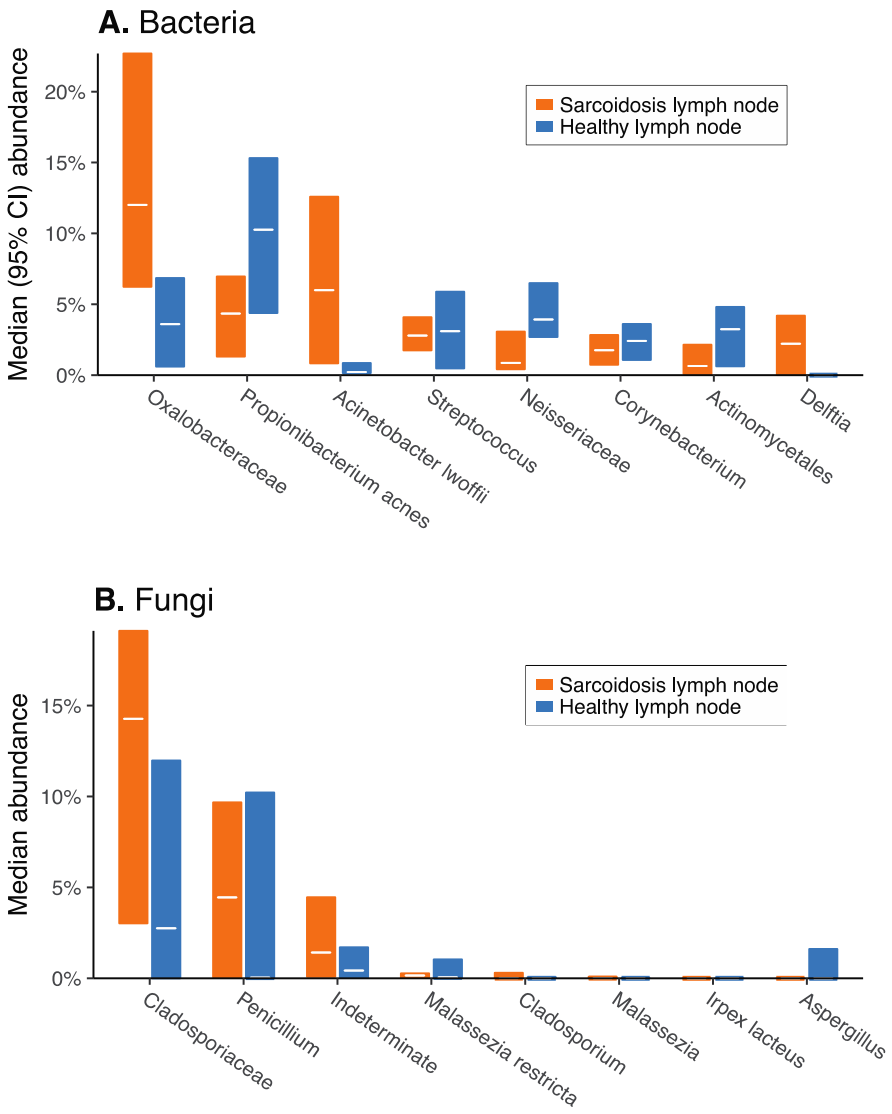


Supp. Figure 2-2. Dominant taxa in tissue samples (Set A).
The dominant bacterial taxa (A) and fungal taxa (B) in sample set A are shown in the form of their median abundance in all samples (white lines) and 95% median confidence intervals (surrounding boxes). Lineages are grouped by their most specific taxonomic rank available. A differential enrichment of Aspergillus fungi is visible in the sarcoidosis samples and environmental controls, and enrichment of Cladosporiaceae is apparent in sarcoidosis lymph node.

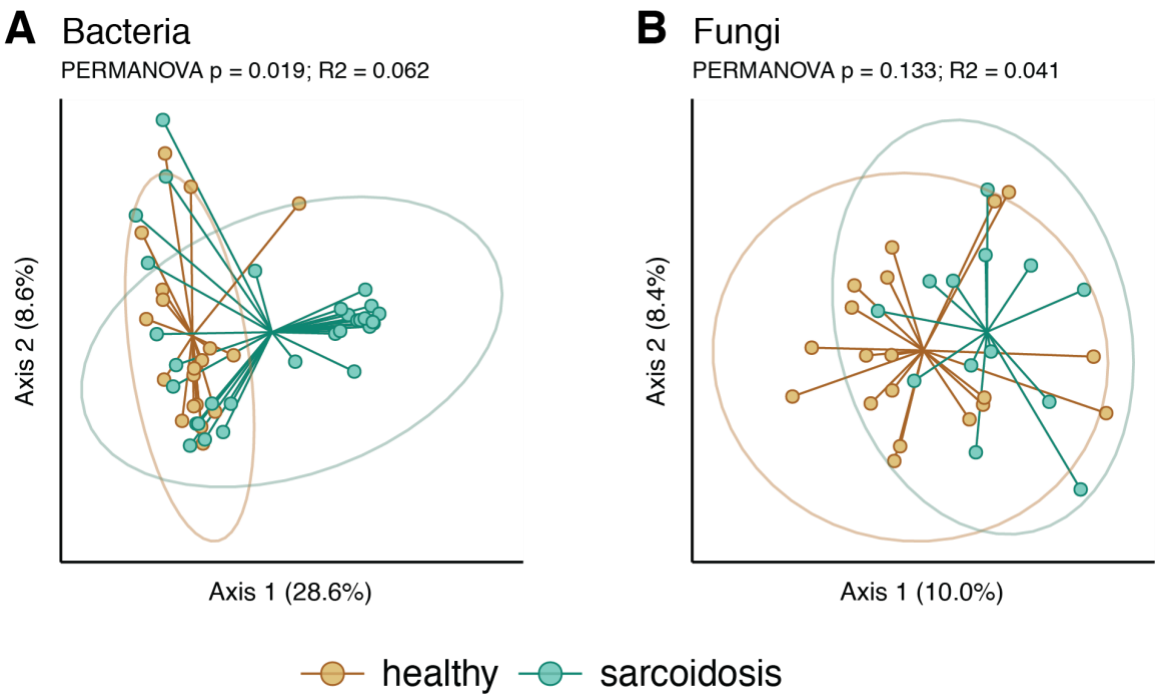


Supp. Figure 2-3. Community differences in tissue samples (Set A).

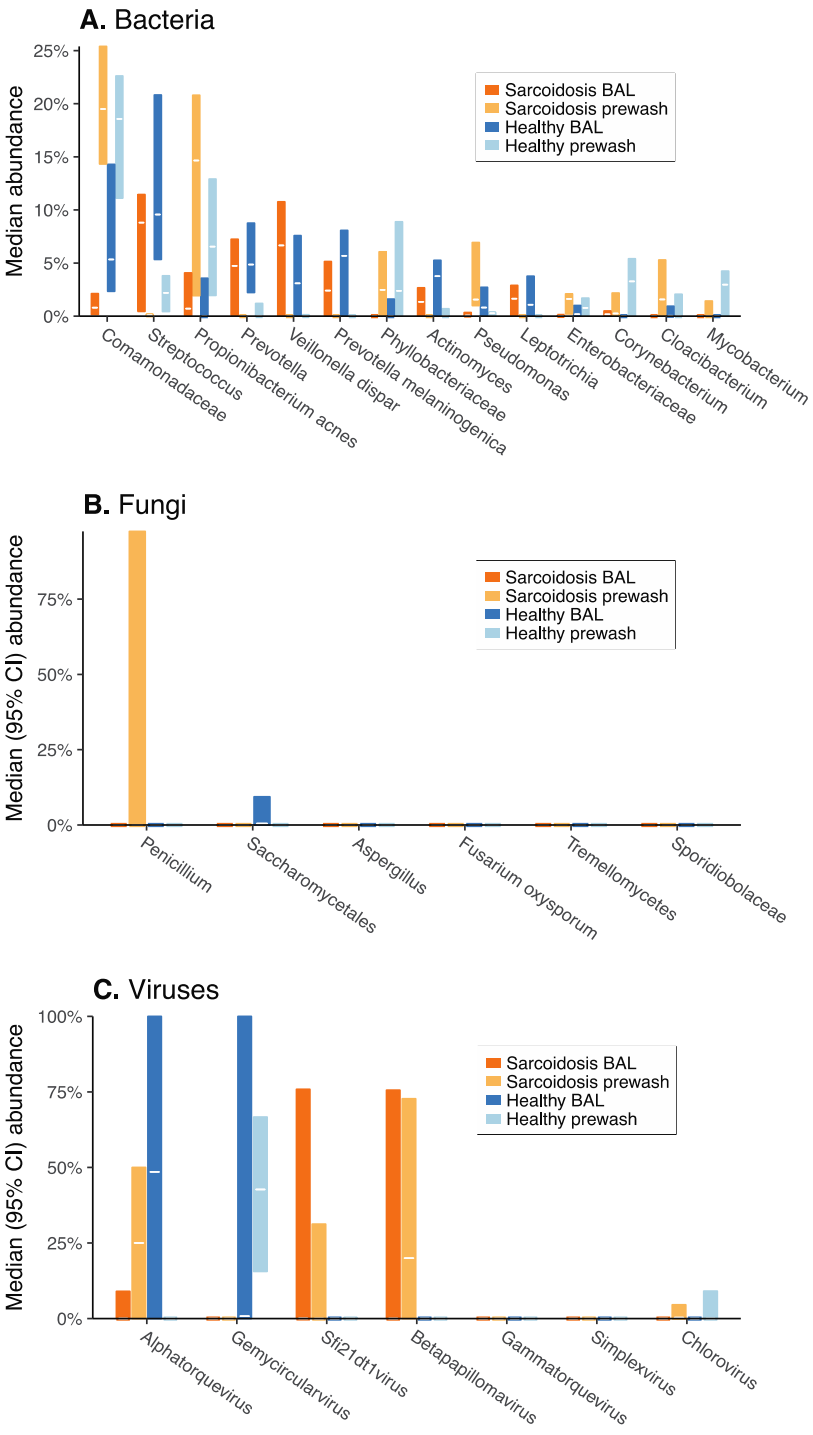
Distances between all pairs of samples were generated using generalized UniFrac (J. Chen et al., 2012; Lozupone, Lladser, Knights, Stombaugh, & Knight, 2011) with an alpha parameter of 0.5. Principle Coordinate Analysis (PCoA) plots show relationships between bacterial communities in lymph nodes (A), fungal communities in lymph nodes (B), bacterial communities in blank paraffin (C), and fungal communities in blank paraffin (D). There were no significant differences in bacterial communities between sarcoid and healthy lymph node samples or their matched paraffin controls. There were significant differences in the fungal communities of sarcoid and healthy lymph node samples ($p < 0.05$), but these differences were also present in the matched paraffin controls ($p < 0.01$).



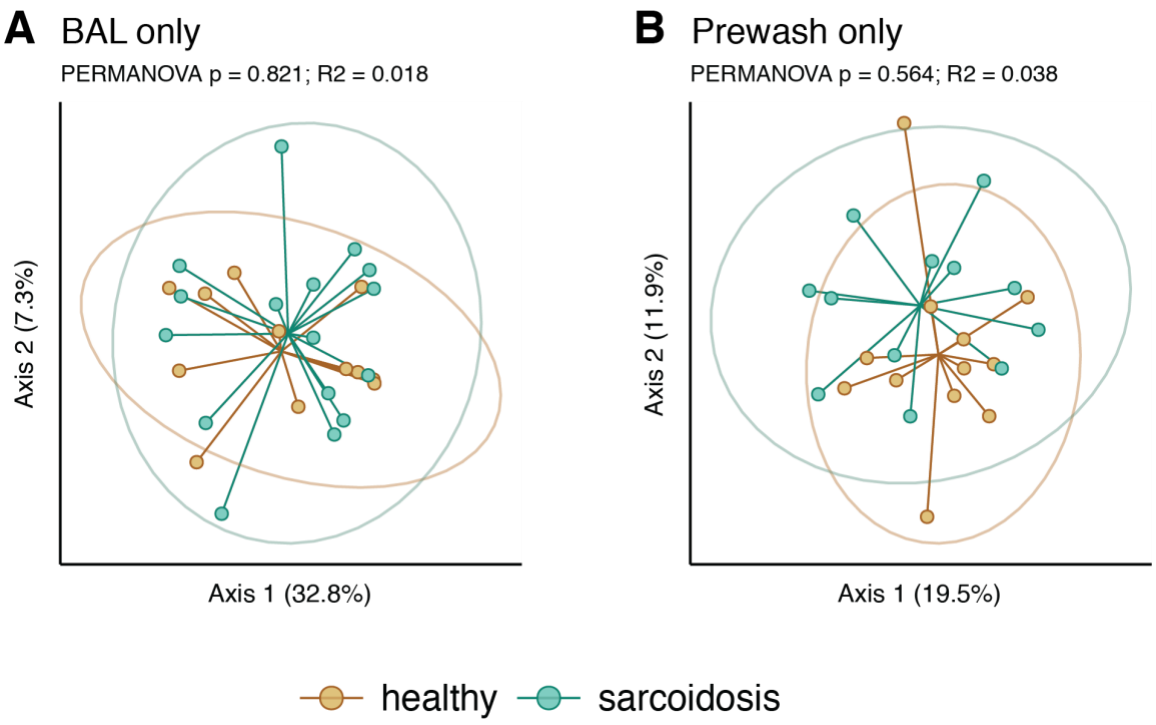
Supp. Figure 2-4. Dominant taxa in tissue samples (Set B).
The dominant bacterial taxa (A) and fungal taxa (B) in sample set B are shown in the form of their median abundance in all samples (white lines) and 95% median confidence intervals (surrounding boxes). Lineages are grouped by their most specific taxonomic rank available.



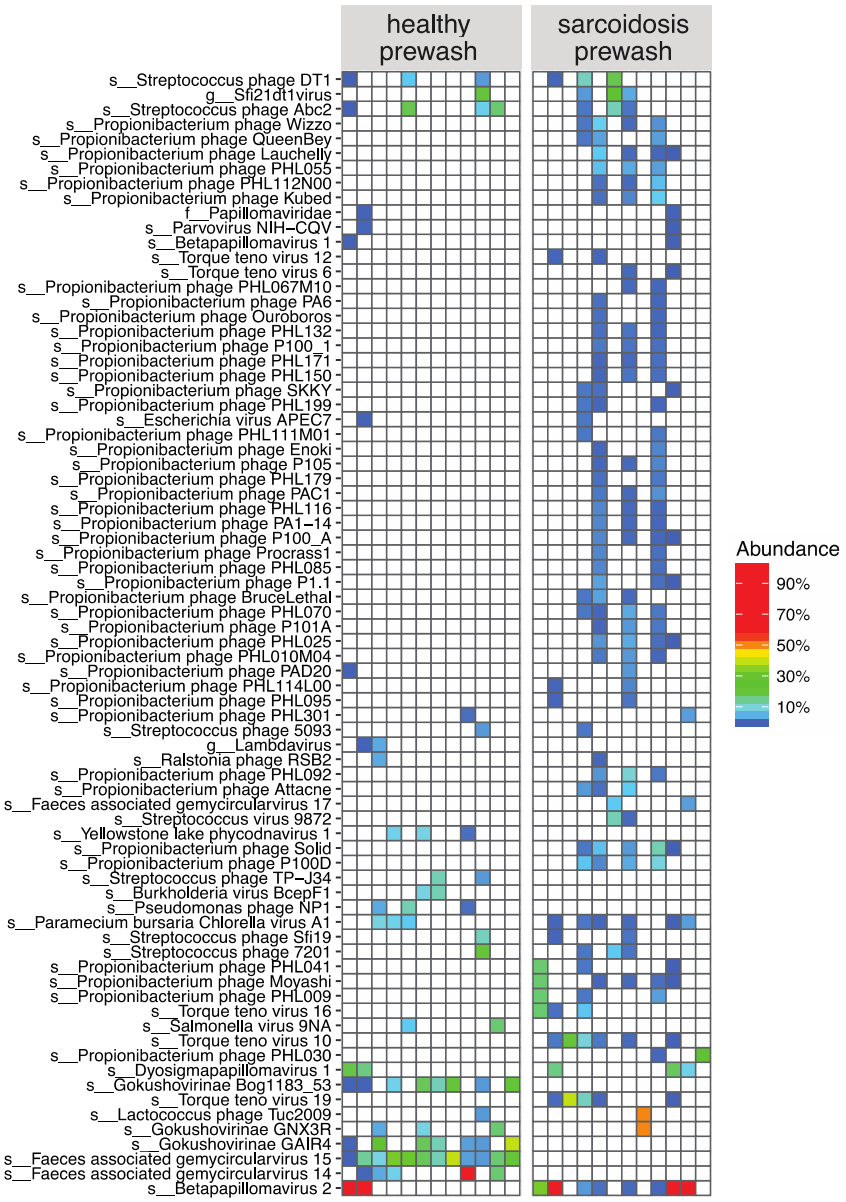
Supp. Figure 2-5. Community differences in tissue samples (Set B).
Distances between all pairs of samples were generated using generalized UniFrac (alpha=0.5). Principle Coordinate Analysis (PCoA) plots showing relationships between bacterial communities (A) and fungal communities (B) in lymph node samples in sample set B. Significant differences were found in the bacterial communities of sarcoidosis and healthy lymph node ($p < 0.05$).



Supp. Figure 2-6. Dominant taxa in BAL samples (Set C).
The dominant bacterial taxa (A), fungal taxa (B) and viruses (C) in sample set C are shown in the form of their median abundance in all samples (white lines) and 95% median confidence intervals (surrounding boxes). Lineages are grouped by their most specific taxonomic rank available.



Supp. Figure 2-7. Community differences in BAL samples (Set C).
Distances between all pairs of samples were generated using generalized UniFrac (alpha=0.5). Principle Coordinate Analysis (PCoA) plots showing relationships between bacterial communities in BAL and prewash samples from set C. No significant differences were found between sarcoidosis and healthy BAL (panel A) or their corresponding prewash (panel B).



Supp. Figure 2-8. Differences in prewash viral populations.
Viral populations differ between bronchoscope prewash samples taken just before bronchoscopy of healthy volunteers (left panel) and sarcoidosis subjects (right panel). Rows indicate the viruses detected, columns are separate samples.

2.7.3. Supplemental Tables

Supp. Table 2-1. Oligonucleotide sequences used in Chapter 2.

	Name	*Sequence*	*Components*
Illumina MiSeq 16S V1-V2 Amplification and Sequencing Primers	Forward PCR amplification primer	5'-AATGATACGGCGACCACCGAGATCTACAC-XXXXXXXXXXXX-TATGGTAATT-GT-AGAGTTTGATCCTGGCTCAG-3'	5' Illumina adapter, 12-base Golay barcode (X's), forward pad*, linker**, 27F forward primer
	Reverse PCR amplification primer	5'-CAAGCAGAAGACGGCATACGAGAT-XXXXXXXXXXXX-AGTCAGTCAG-CC-TGCTGCCTCCCGTAGGAGT-3'	Reverse complement of 3' Illumina adapter, 12-base Golay barcode (X's), reverse pad, linker, 338R reverse primer
	Forward sequencing primer	5'-TATGGTAATT-GT-AGAGTTTGATCCTGGCTCAG-3'	Forward pad, linker, 27F forward primer
	Reverse sequencing primer	5'-AGTCAGTCAG-CC-TGCTGCCTCCCGTAGGAGT-3'	Reverse pad, linker, 338R reverse primer
Illumina MiSeq ITS1F and ITS2 Amplification and Sequencing Primers	Forward PCR amplification primer	5'-AATGATACGGCGACCACCGAGATCTACAC-XXXXXXXXXXXX-TGCGGCCTGC-GT-CTTGGTCATTTAGAGGAAGTAA-3'	5' Illumina adapter, 12-base Golay barcode (X's), forward pad*, linker**, ITS1F primer
	Reverse PCR amplification primer	5'-CAAGCAGAAGACGGCATACGAGAT-XXXXXXXXXXXX-AGTCAGTCAG-CC-GCTGCGTTCTTCATCGATGC-3'	Reverse complement of 3' Illumina adapter, 12-base Golay barcode (X's), reverse pad, linker, ITS2 primer
	Forward sequencing primer	5'-TGCGGCCTGC-GT-CTTGGTCATTTAGAGGAAGTAA-3'	Forward pad, linker, ITS1F primer
	Reverse sequencing primer	5'-AGTCAGTCAG-CC-GCTGCGTTCTTCATCGATGC-3'	Reverse pad, linker, ITS2 primer

Chapter 3. Swga: A primer design toolkit for selective whole genome amplification

The contents of this chapter have been previously published in:

> Clarke, E. L., Sundararaman, S. A., Seifert, S. N., Bushman, F. D., Hahn, B. H., & Brisson, D. (2017). swga: a primer design toolkit for selective whole genome amplification. *Bioinformatics, 33*(14), 2071-2077. doi:10.1093/bioinformatics/btx118

3.1. Abstract

Motivation: Population genomic analyses are often hindered by difficulties in obtaining sufficient numbers of genomes for analysis by DNA sequencing. Selective whole-genome amplification (SWGA) provides an efficient approach to amplify microbial genomes from complex backgrounds for sequence acquisition. However, the process of designing sets of primers for this method has many degrees of freedom and would benefit from an automated process to evaluate the vast number of potential primer sets.

Results: Here, we present *swga*, a program that identifies primer sets for SWGA and evaluates them for efficiency and selectivity. We used *swga* to design and test primer sets for the selective amplification of *Wolbachia pipientis* genomic DNA from infected *Drosophila melanogaster* and *Mycobacterium tuberculosis* from human blood. We identify primer sets that successfully amplify each against their backgrounds and describe a general method for using *swga* for arbitrary targets. In addition, we describe characteristics of primer sets that correlate with successful amplification, and present guidelines for implementation of SWGA to detect new targets.

Availability and Implementation: Source code and documentation are freely available on https://www.github.com/eclarke/swga. The program is implemented in Python and C and licensed under the GNU Public License.

3.2. Introduction

Selective whole-genome amplification (SWGA) provides a means of obtaining sufficient numbers of genomes from a target organism to perform whole-genome sequence analysis, even in the presence of overwhelming DNA from other organisms (Leichty & Brisson, 2014). Difficulties in isolating a target of interest are common in microbial population genomics, which requires acquiring adequate genomic DNA from a target while limiting the amount of non-target DNA (Mardis, 2008). Often, the genomes of interest represent only a fraction of a percent of the total nucleic acids in a sample, and so direct sequencing is inefficient and expensive. Laboratory culture of the target microbe is the traditional solution, but many microbes replicate poorly or not at all in in vitro conditions (Amann et al., 1990; Ghazanfar et al., 2010; Schmeisser et al., 2007).

SWGA allows sequence acquisition without culture of the target organism or extensive purification of target DNA. It achieves this by preferentially amplifying the target genome using a set of selective primers and phi29 polymerase-based multiple displacement amplification (MDA) (Dean et al., 2002; Leichty & Brisson, 2014). Since its introduction, this method has been used to study *Wolbachia pipientis* in *Drosophila melanogaster* (Leichty & Brisson, 2014), and to understand the evolution and drug resistance of *Plasmodium falciparum* (Guggisberg et al., 2016; Oyola et al., 2016;

Sundararaman et al., 2016) and *Plasmodium vivax* (Cowell et al., 2017). Further applications of SWGA to population genomics may help reconstruct epidemic transmission patterns, characterize patterns of inter-host viral transmission, detect escape from antimicrobial agents, and delineate the evolutionary dynamics of immune escape (Hume, Lyons, & Day, 2003; Luikart, England, Tallmon, Jordan, & Taberlet, 2003; Martínez et al., 2012; Nelson et al., 2008; Nunes et al., 2012; Stack, Murcia, Grenfell, Wood, & Holmes, 2012).

Implementation of SWGA has been complicated by the difficulty in identifying an effective set of selective primers, as there are many constraints and degrees of freedom in the composition of potential primer sets. These primers must reflect DNA sequence motifs common in the target genome but rare in the background DNA. They also must have binding sites sufficiently near each other to enable the branching and displacement actions of the phi29 polymerase that are essential for MDA. A previously published method used a set of Perl scripts (Leichty & Brisson, 2014) to identify primers with the highest ratio of binding frequencies in the target genome versus the background DNA. However, choosing a set by the above method is suboptimal: for one, the primers may form heterodimers with each other or homodimers with themselves; they may be individually selective but in aggregate bind too frequently to the background DNA; or, they may bind to the target's telomeric or mitochondrial DNA, and not be sufficiently evenly distributed across the genome. There are aspects of the primer sets that have an unknown effect on the efficiency of the reaction, including the annealing and melting temperature of the primer sequences, the evenness of the binding sites across the target

genome, and the density of binding sites. The Perl scripts mentioned above are unable to evaluate many of these criteria, requiring extensive manual effort and trial-and-error to create workable designs.

Here we present *swga*, a program that identifies selective primer sets for a given target genome and background. *swga* evaluates all potential primer sequences and forms sets of valid primers that meet the above criteria. It automatically calculates a variety of metrics for each set that potentially affect the efficacy and selectivity of the reaction. These sets are then are ranked and presented to the user, enabling the selection of primer sets most likely to succeed. Nearly all operating parameters of the program are user-specifiable but initialized with reasonable defaults based on the target and background genomes selected, reducing the work needed to get started.

We demonstrate the use of *swga* to design primer sets and test them on two biological systems: *Wolbachia pipientis* from infected *Drosophila melanogaster*, and *Mycobacterium tuberculosis* DNA spiked into human blood. For each system, we designed multiple primer sets to explore the effect of various aspects of the primer sets on reaction efficacy, such as primer melting temperature, binding density on the target genome, and the evenness of binding sites. These experimental results clarify the relative importance of each and allow us to describe an effective workflow for using *swga*.

3.3. Methods

3.3.1. Program overview

The *swga* program can be divided into four modules (Fig. 1). The user starts by defining the target and background sequences using **swga init**. At this point, a set of sequences can be supplied that define a priori where primers should not bind, such as a mitochondrial genome or plasmids (the 'exclusionary sequences'). The **swga count** command then uses DSK (Rizk, Lavenier, & Chikhi, 2013) to identify all nucleotide sequences in the size range specified by parameters min_size and max_size that exist in the target genome and do not exist in the exclusionary sequences (if provided). These primers are used to populate a local SQLite database for later retrieval. The selectivity of these primers is determined by their frequency in the target genome versus the background DNA, so **swga count** saves the frequency that each primer appears in the target and background as well. Primers that appear extremely rarely in the target and overly frequently in the background (as defined by user-editable parameters, with defaults set by **swga init**), are not saved to help speed up downstream steps. Additionally, primers that would form internal hairpins or homodimers with themselves are omitted.

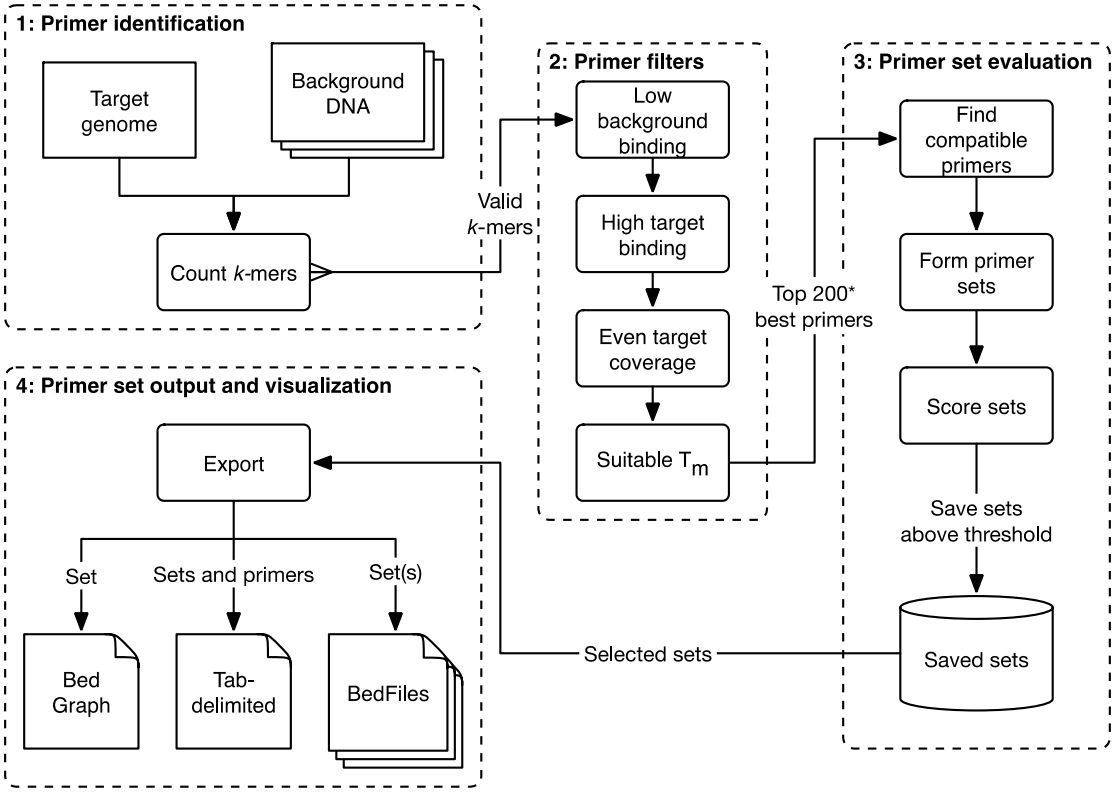


Figure 3-1. An overview of the swga workflow.
An overview of the swga workflow. The program begins by counting all nucleotide sequences of length k (k-mer) in both the target and background genomes for a given range of k (e.g. 8–12 bp). The k-mers are then filtered by criteria that include the binding frequencies in the background and target genome, their melting temperatures, and the likelihood of hairpin or homodimer formation. The best k-mers are then used to form compatible sets, in which no k-mer would likely form a heteroduplex with any other in the set. These sets are then evaluated for multiple criteria including binding frequencies and evenness. The results can be exported into common formats for downstream use and visualization.

3.3.2. Primer filtering

The command **swga filter** ranks and filters potential primers by their melting temperature, selectivity, and evenness of binding in the target genome. First, primers that bind too sparsely to the target genome (lower than parameter min_fg_bind) or too frequently to the background (max_bg_bind) are removed. Next, the melting temperature is approximated using nearest-neighbor thermodynamics (Allawi & SantaLucia, 1997) with corrections for mono- and divalent cations. Primers with melting temperatures outside the range defined by min_tm and max_tm are removed. The evenness of binding then is calculated by finding the Gini index (Gini, 1912) of the distances between each primer binding site on the target. The Gini index varies between 1 and 0, where 1 represents extremely uneven and 0 represents perfectly even. A primer with a low Gini index has binding sites that are each separated by similar distances, whereas a primer with a high Gini index may reflect one where many of the primer binding sites are clumped together (e.g. on tandem repeat regions). Primers with Gini indices higher than max_gini are removed. Finally, primers are ranked by the ratio of target binding frequency to background binding frequency and those primers with the highest ratio are identified for downstream use (by default, this identifies the top 200 primers, and is modifiable via the max_primers parameter). The thresholds for each filter are user-editable, and the **swga filter** command caches results so that it can be quickly re-run to explore different results.

3.3.2.1. Primer set evaluation

The **swga find_sets** command is then used to find sets of compatible primers from the ones identified in the last step of **swga filter**. Brute force evaluation of all primer sets is computationally infeasible: given *n* primers and a set size of *k*, the total number of possible sets is (*n* choose *k*). With the default parameters of n=200 and k=2−7, there are over 2.4×10^{16} possible sets. Fortunately, not all of these sets are usable for *swga*. A pair of primers are incompatible if they form heterodimers (calculated by the number of consecutive complimentary bases), or if one primer is a subsequence of another. **swga find_sets** calculates the pairwise compatibility of all selected primers and stores the results as a graph. In this graph, primers are vertices and compatible primers are connected with edges. The problem of finding compatible sets then reduces to a problem of finding sets of vertices in the graph that are all interconnected (a 'clique' in graph theory). *swga* also stores the average distance between binding sites on the background as a 'weight' on each vertex. This allows the program to prioritize cliques that have higher total weights, representing sets of primers that bind infrequently to the background.

To find these cliques, the **swga find_sets** command uses a modified version of the program *cliquer* (Niskanen & Östergård, 2003). The branch-and-bound algorithm in *cliquer* is a computationally efficient way of finding cliques in a graph. We have extended the algorithm to find only cliques that meet certain criteria. By specifying the desired criteria *a priori* the algorithm can skip sets that do not meet the requirements and save computation time. These criteria include the minimum distance between binding sites in the background (min_bg_bind_dist) and maximum distance between binding sites

on the target (max_fg_bind_dist). In addition, the algorithm can explore a range of set sizes (min_size and max_size) in order to find valid sets. By specifying a broad range of set sizes, the algorithm is able to find sets with a broad range of characteristics independent of the number of primers.

Primer sets that meet these criteria are further evaluated on metrics including the average and maximum distance between primer binding sites on the target genome and the Gini index of all binding sites in the set. These sets and their accompanying metrics are then saved. Even with the above optimizations, the number of valid sets can be quite large. For this reason, **swga find_sets** can be safely stopped after evaluating and storing a sufficient number of sets. In our usage, we generally stop after 1–5 million sets have been saved.

3.3.2.2. Primer set output and visualization

The saved primer sets can be explored and exported using **swga export**. This command allows the user to order the sets by any of the evaluated metrics, export all or some of the sets of interest to Excel-compatible formats, or export a set to a BedGraph or BedFile format for visualization in a genome browser (Kent et al., 2002)

3.3.3. Empirical primer set testing

To evaluate *swga*, we used it to design primer sets for amplification of *W. pipientis* DNA against a background of *D. melanogaster* and of *M. tuberculosis* against a background of H. sapiens. We evaluated primer sets on their ability to selectively and evenly amplify the target genome.

3.3.3.1. Designing primer sets for W.Pipientis

We created four primer sets for *W. pipientis* against *D. melanogaster*, varying each by melting temperature range, selectivity, and evenness of binding sites on the target genome. We first initialized *swga* on the *W. pipientis* genome with D. melanogaster as the background, and ran **swga count** to store all potential primers.

For the first two sets, we used **swga filter** with the 'standard' temperature range established in Leichty and Brisson (2014), and default in *swga*, of 15–45°C. This range we named Tm Low, or TmL. After running **swga find_sets** and storing 1 million sets, we used **swga export** to output the set with the lowest target to background binding distance ratio, which we called Set TmL/Selective. We then used **swga export** again to output the set with the lowest Gini index, which we called Set TmL/Even.

The next two sets were designed with a higher melting temperature range. We re-ran **swga filter** with a Tm range of 35–55°C, which we named Tm High, or TmH. As above, we then re-ran **swga find_sets** on the new primers and chose the most selective and most even sets from the results. These are called TmH/Selective and TmH/Even, respectively. The complete parameter listing is included in Section §3.7.3. The primers belonging to each set are given in Supp. Table 3-1.

3.3.3.2. Designing primer sets for M. tuberculosis

We created ten primer sets for *M. tuberculosis* using *swga*. Our target genome was *M. tuberculosis* strain H37Rv (NC_000962.3) and our background was the human genome, version GRCh38. For this system, we ran **swga filter** with a temperature range constant

at 15–45°C, and imposed a maximum per-primer Gini index of 0.6. We stopped **swga find_sets** after storing five million sets and exported all of them to CSV format using **swga export**. The sets were filtered to only sets with mean distance between target binding sites <5000 bases. We selected ten sets with the most extreme combinations of mean target binding distance and evenness (via the metrics fg_dist_mean and fg_dist_gini, respectively). These sets we named Mtb1 through Mtb10. The distribution of these sets in the pool is visualized in Supp. Figure 3-1. In addition, we selected from the original five million the set with the highest Gini index (most uneven) and highest mean target binding distance as negative comparisons, named MtbUneven and MtbSparse, respectively. The full parameter listing is included in Section §3.7.3.2. The primers belonging to each set are given in Supp. Table 3-2.

3.3.4. Selective whole-genome amplification and sequencing

The *Wolbachia*-specific primer sets were tested on pooled genomic DNA extracted from 10 *Wolbachia*-infected *D. melanogaster* (strain Dmel\w^{118}). Pooling was performed to eliminate inter-fly variability in *Wolbachia* infection levels, and each primer set was tested in triplicate using 40 ng of input DNA per reaction, except as noted for additional tests of the TmL/Even *Wolbachia* primer set. For consistency with the approach used in Leichty and Brisson (2014), the pooled genomic extract was digested with *NarI* (NEB, New England Biolabs, Inc., Ipswich, MA, USA) at 37°C for 30 minutes, in order to suppress mitochondrial amplification. This step is likely unnecessary in the general case because *swga* includes an option to omit mitochondrial sequences from primer formation.

Mycobacterium primer sets were tested on purified *M. tuberculosis* DNA (strain H37Rv, ATCC 27294D-2), diluted to 1% in human genomic DNA extracted from cultured CD4+ T cells. Primer sets were tested in triplicate.

Selective whole-genome amplification was performed as previously described (Sundararaman et al., 2016), with slight modifications. Reactions were performed in a volume of 50 μL using input DNA, 3.5 mM total of SWGA primers, 1× phi29 buffer (New England Biolabs), 1 mM dNTPs and 30 units phi29 polymerase (New England Biolabs). Amplification conditions included a 1 h ramp-down step (35–30°C), followed by a 16 h amplification step at 30°C. Phi29 was then denatured for 10 min at 65°C.

Amplified samples were purified using AmpureXP beads (Beckman Coulter), prepared for Illumina sequencing as described in (Kryazhimskiy, Rice, Jerison, & Desai, 2014), and sequenced on an Illumina MiSeq (150 bp, paired end). We also sequenced the unamplified pool to establish a baseline for amplification efficiency. Illumina-specific adapter and primer sequences were removed from the reads using *cutadapt* (Martin, 2011). In both systems, reads were first aligned to the background (*D. melanogaster* or human) using *smalt* (Ponstingl & Ning, 2010). Unmapped reads were then mapped to the target genome (*W. pipientis* or *M. tuberculosis*, respectively), also using *smalt*. Analysis of sequence coverage of the target genome and sequencing rarefaction analyses were performed using R (R Core Team, 2017). All code used in the analysis and to generate the figures is available online at https://github.com/eclarke/swga_paper.

3.4. Results

We used *swga* to design four primer sets for amplifying *Wolbachia* against a background of *D. melanogaster*, which tested the effect of melting temperature ranges, selectivity, and evenness. We designed twelve primer sets for amplifying *M. tuberculosis* against a background of human DNA with varying primer binding evenness and density on the target. Ten sets tested were various combinations of high density and evenness. Two, for comparison, were the most uneven and most sparse. For *M. tuberculosis*, we compared amplification using random hexamers (e.g. standard MDA) to the *swga*-designed primer sets.

3.4.1. Evaluation of primer sets for W.pipientis

The four primer sets for *W. pipientis* were designed with two different temperature ranges (TmL: 15–45°C, TmH: 35–55°C). From the sets identified in each temperature range, we chose the set with the highest selectivity, defined by the lowest target to background binding distance ratio (TmL/Selective and TmH/Selective). We also chose the sets with the most even distribution of binding sites (TmL/Even and TmH/Even). As a control, we included the primer set from Leichty and Brisson (2014). The composition and metrics for each of these five sets is shown in Table 3-1.

Table 3-1. Primer sets for Wolbachia.

	Ratio	*# Primers*	*Gini*	*Mean target dist*	*Mean bg. dist*
T_mL/Selective	0.0544	9	0.654	5.33E+03	9.78E+04
T_mL/Even[a]	0.1050	7	0.537	6.85E+03	6.53E+04
T_mH/Even	0.0075	2	0.537	1.31E+04	1.73E+06
T_mH/Selective	0.0005	2	0.66	1.21E+04	2.43E+07
Leichty 2014	0.0163	2	0.712	5.31E+03	3.25E+05

Characteristics of primer sets chosen for selective whole-genome amplification of Wolbachia from infected Drosophila DNA. 'Ratio' is the ratio of average distance between binding distances in the target and background. [a]The set that most effectively amplified Wolbachia.

The pooled genomic DNA contained 4.7% *W. pipientis* DNA, as determined by sequencing of the unamplified control. We recovered ~200 Mbp of sequence for each amplicon. The proportion of sequencing reads that were derived from *W. pipientis* was at least 2.5 times greater in all amplified samples than the sequencing reads from the unamplified genomic extract (Supp. Figure 3-2). We found that the primer sets with the higher melting temperatures (TmH/Selective and TmH/Even) yielded more *Wolbachia* reads as a total percentage, with some replicates as high as 77.8%. However, these primer sets failed to reach 10× coverage on even 10% of the *W. pipientis* genome (Figure 3-2). This was most likely due to uneven amplification of the target genome, as shown in Supp. Figure 3-3.

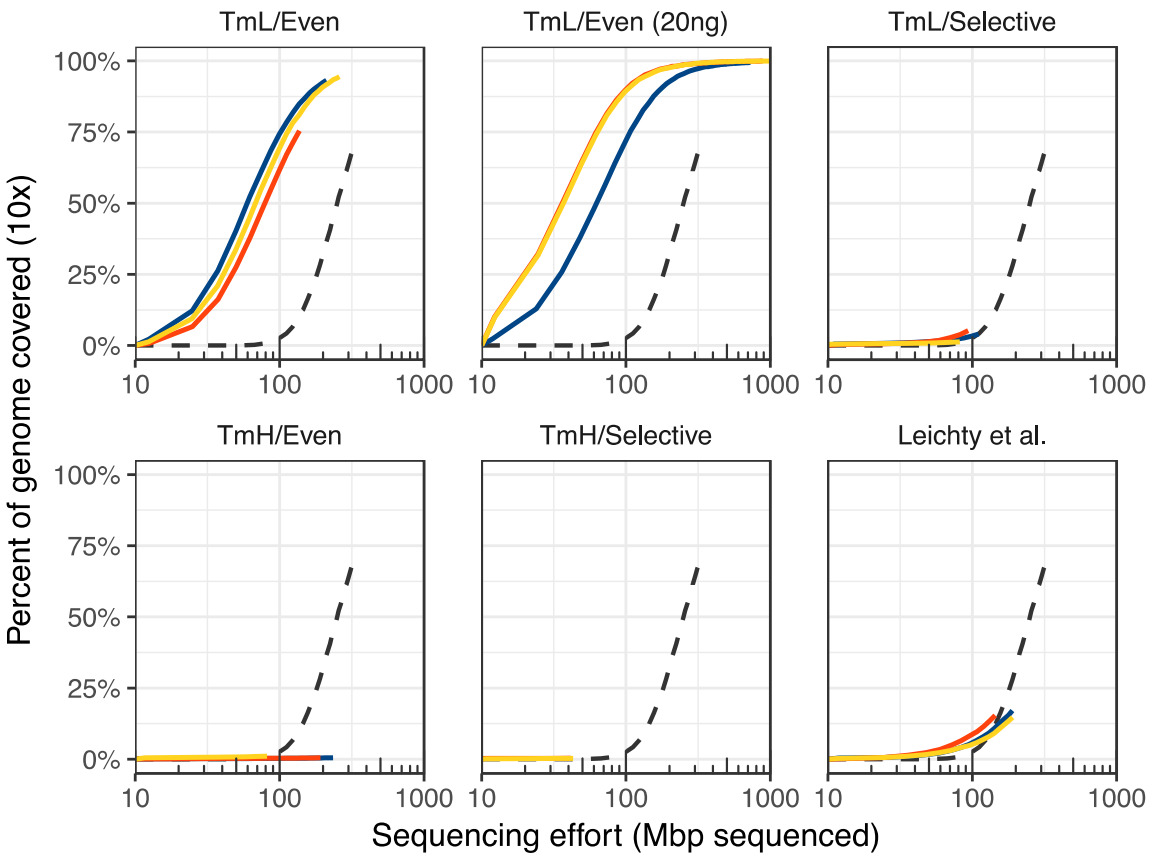


Figure 3-2. Sequencing effort required to cover Wolbachia genome.
Selective whole genome amplification reduces the sequencing effort necessary to achieve at least ×10 coverage across the W. pipientis genome. Each color represents an individual technical replicate; dashed lines represent the unamplified control. Lines above the unamplified control represent better sequencing efficiency in that they yielded greater coverage of the target genome with less sequencing effort. Sequencing 100 million bases from unamplified genomic DNA extracted from 10 flies resulted in 10-fold or greater sequencing coverage in only 2.8% of the W. pipientis genome. In contrast, the T_mL/Even primer set resulted in 10-fold or greater coverage of 60–75% of the W. pipientisgenome with similar sequencing effort. This fraction was increased further to 72–91% when the T_mL/Even primer set was used to amplify W. pipientis from 20 ng (rather than 40 ng) of total fly extract DNA (empirically, using lower total starting DNA can yield higher relative amplification when using phi29). The T_mL/Selective primer set and the manually chosen set (Leichty & Brisson, 2014) improved W. pipientis sequence coverage relative to the unamplified sample. However, both of these sets failed to improve sequencing efficiency due an unevenness of coverage. The high T_m sets enriched only small portions of the genome and thus did not improve the genome coverage relative to the control.

In contrast, the sets designed with the standard, lower melting temperature range (TmL) yielded more even coverage across the genome (Supp. Figure 3-3). The TmL/Even primer set, selected for having the most even distribution of primer sites across the

Wolbachia genome, gave high, even coverage across the target (Figure 3-3; Supp. Figure 3-3). Moreover, the TmL/Even set reduced the sequencing effort required to achieve 10× coverage across 90% of the genome by 10-fold relative to the unamplified control (Figure 3-2), extrapolating from the still-rising unamplified control's rarefaction curve. While the final two sets—TmL/Selective and the Leichty set—provided more even coverage of the genome than the TmH sets, they ultimately did not outperform the unamplified control. The previously-published primer set from Leichty and Brisson (2014) yielded low total amplification efficiency (12.1–27.7%) and uneven coverage, while the TmL/Selective set had high amplification efficiency (50–60%) but similarly uneven coverage.

We had originally expected that high numbers of primer binding sites in local regions of the genome would provide better coverage of that region. This was not seen in any of the sets tested (Supp. Figure 3-4). In each of the five sets tested, we did not detect a correlation between the number of primer binding sites and coverage. However, in primer sets with an overall higher density of binding sites on the target (as measured by a low average distance between binding sites), we had generally higher coverage across more of the *Wolbachia* genome (compare Figure 3-2 and Table 3-1).

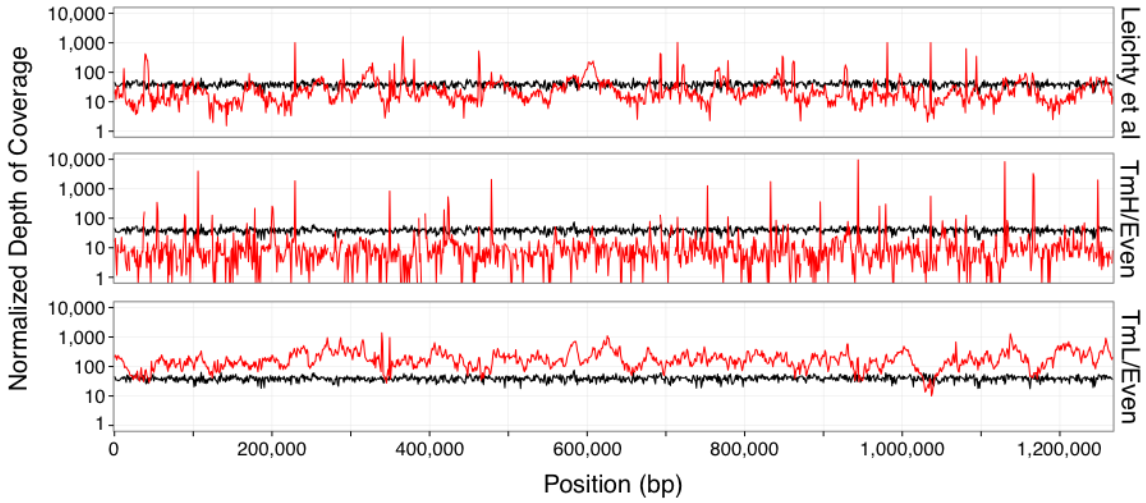


Figure 3-3. Genome coverage by primer sets.
Sequencing coverage of two swga-chosen sets, and the set from Leichty and Brisson (2014), across the W. pipientis genome. The depth of sequencing coverage per 1 Mb of sequencing effort (1 Mb * coverage depth/total bp sequenced) is shown for representative replicates of T_mH/Even and T_mL/Even (red lines) relative to the unamplified control (black lines). SWGA using the T_mL/Even primer set improves depth of coverage across the majority of the W. pipientis genome by 10- to 100-fold, relative to the unamplified control. SWGA using the Leichty and Brisson (2014) (top panel) or T_mH/Even (middle panel) sets also improve depth of coverage but over smaller regions of the genome, with the T_mH/Even set resulting in high but localized amplification. Depth of coverage plots for all primer sets and replicates are shown in Supp. Figure 3-3.

In summary, the primer set with the lowest Gini index and standard melting temperature (TmL/Even) was the best at selectively and evenly amplifying *Wolbachia*. While other sets provided a higher percentage of *Wolbachia* DNA (Supp. Figure 3-3), the overall coverage of these sets was low and amplification mostly occurred in specific regions (Figure 3-3). This suggests that evenness of primer binding sites on the target is a major factor in the efficacy of the primer set.

3.4.2. Evaluation of primer sets for M. tuberculosis

For *M. tuberculosis*, we restricted the primer pool to only those with a low Gini index (<0.6). We let the program identify five million primer sets and then selected only those

sets whose mean distance between binding sites on the *M. tuberculosis* genome was <5 kb. From the resulting pool of primer sets, we selected ten sets with the most extreme combinations of primer set binding evenness and density to test the contributions of each. These ten will be referred to as our positive tests (Mtb1-10), and the distribution of these points on the total pool of sets is shown in Supp. Figure 3-1. We also selected the least selective set and the most uneven set from the five million set pool as negative controls (MtbSparse and MtbUneven, respectively). The composition and metrics for each of these 12 sets is shown in Table 3-2.

Table 3-2. Primer sets for M. tuberculosis.

	Ratio	*# Primers*	*Gini*	*Mean target dist.*	*Mean bg. dist.*
Mtb6[a]	0.0057	7	0.501	1.95E+03	3.41E+05
Mtb9[a]	0.0058	7	0.538	1.78E+03	3.05E+05
Mtb4[a]	0.0062	7	0.512	1.88E+03	3.04E+05
Mtb8[a]	0.0062	7	0.533	1.80E+03	2.92E+05
Mtb7	0.0066	7	0.499	2.03E+03	3.09E+05
Mtb2	0.0095	7	0.484	3.29E+03	3.45E+05
Mtb5	0.0155	6	0.480	5.00E+03	3.22E+05
Mtb1	0.0171	7	0.476	4.97E+03	2.90E+05
Mtb3	0.0172	7	0.478	4.99E+03	2.90E+05
Mtb10	0.0181	7	0.479	4.29E+03	2.37E+05
MtbUneven	0.0140	2	0.623	1.14E+04	8.10E+05
MtbSparse	0.0387	3	0.505	2.60E+04	6.71E+05

Characteristics of primer sets chosen for selective whole-genome amplification of M. tuberculosis from human DNA, ordered by ratio. 'Ratio' indicates the ratio between the average distance between binding distances in the target and background. Primer sequences are listed in Supp. Table 3-2. [a]The sets that most effectively amplified Mycobacterium using SWGA.

The four sets with the lowest mean binding distance (sets Mtb4, Mtb6, Mtb8 and Mtb9) on the *M. tuberculosis* genome performed better than the unamplified controls, six other positive tests, both negative tests and the random hexamers (Figure 3-4 and Table 3-2).

These sets reached 1× coverage across 38–60% of the *M. tuberculosis* genome with 200 megabases of sequence, while the remaining six positive tests did not perform better than the negative controls (Figure 3-4; Supp. Figure 3-5). These four sets yielded higher coverage across most of the *Mycobacterium* genome than the unamplified controls, while the remaining sets either only amplified certain regions or did no better than unamplified (Supp. Figure 3-6). Deeper sequencing of these four sets' amplicons showed that the sets reached 10× coverage over 29–50% of the target by 1.5 Gbp of sequencing effort, with the unamplified controls only reaching 10× coverage on 2.5% of the target for the same sequencing effort (Figure 3-5).

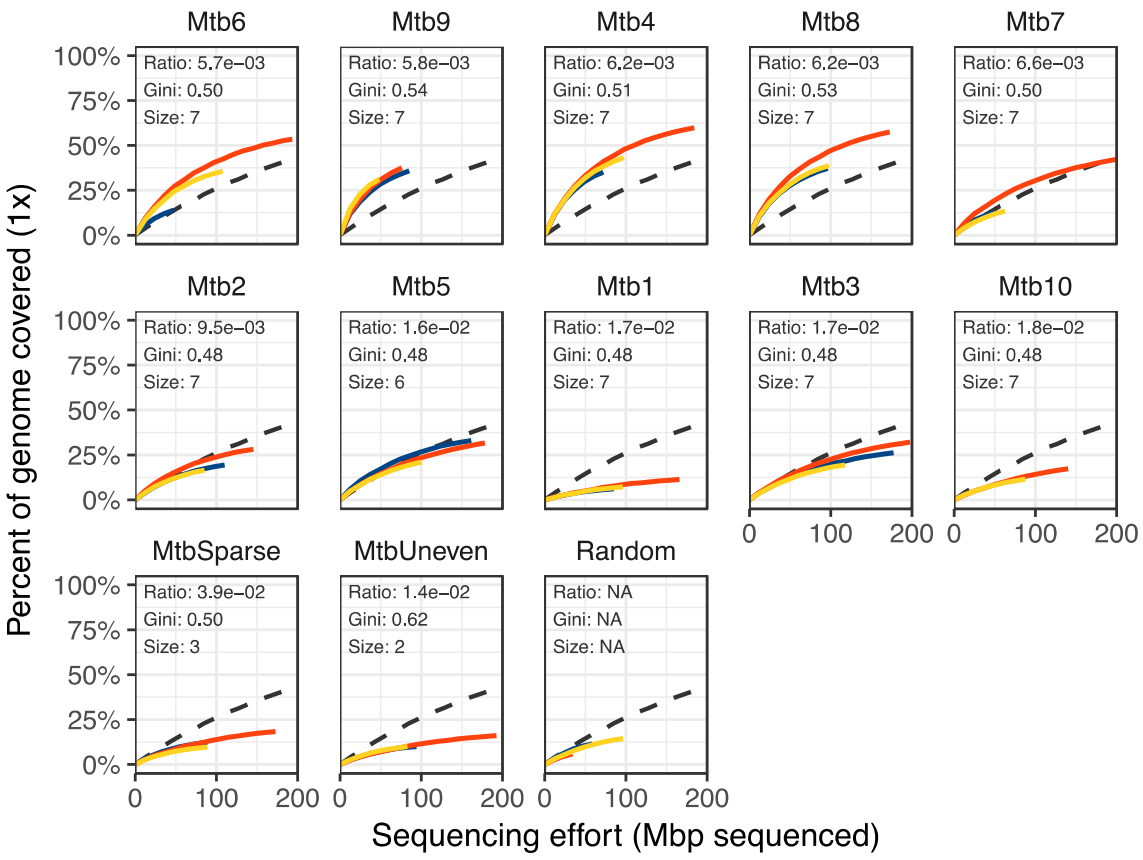


Figure 3-4. Genome coverage using a range of set designs.
Selective amplification of Mycobacterium using swga-designed sets that prioritized primer-level evenness and set-level binding density and selectivity. Curves indicate the percent of the target covered at 1X depth. Sets are ordered by the ratio of average distance between primer binding sites on the target to average binding distance on the background. The coloring indicates individual replicates, and the black dashed line indicates the unamplified control. The sets with the lowest ratios returned greater coverage of the target genome compared to unamplified controls than those with higher ratios, as shown by the rarefaction curves of these sets being higher than the dashed lines.

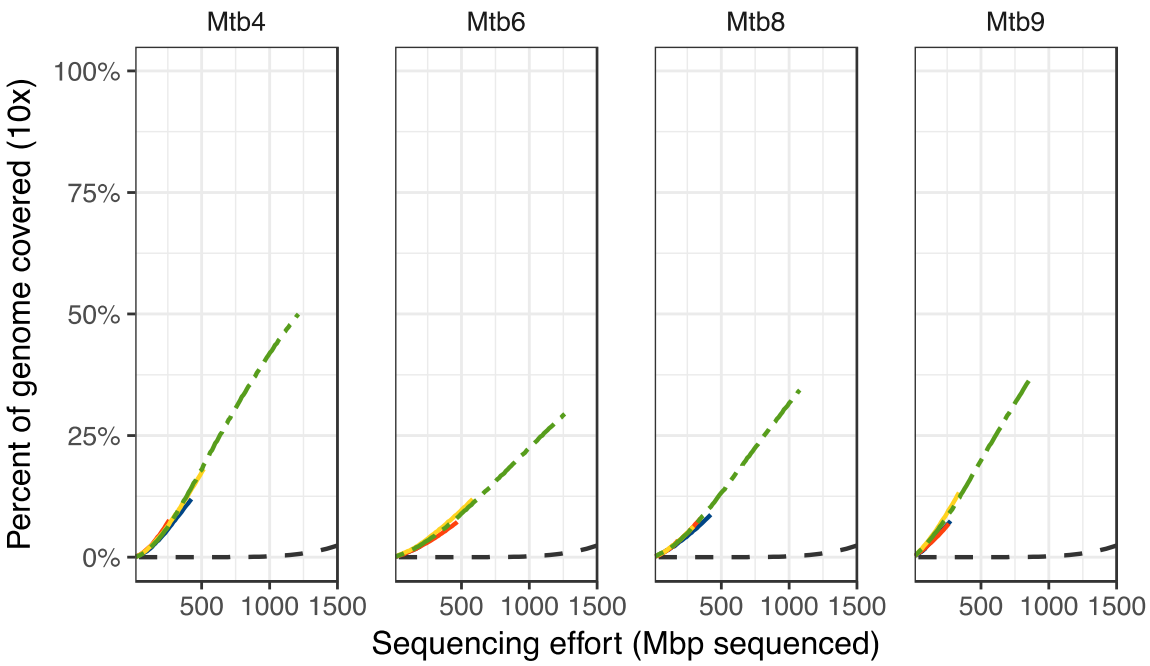


Figure 3-5. Deeper sequencing of M. tuberculosis.
Deeper sequencing of four primer sets yields greater coverage of M. tuberculosis genome. The colored lines indicate individual replicates and the green dashed line is the pooled total. All four sets yield ~10-fold increases in efficiency over the unamplified samples (black dashed line). The primer sets reach ×10 coverage on between 28 and 50% of the target genome while the unamplified controls were at < 2.5% ×10 coverage with 1.5 Gbp of sequencing.

For *Mycobacterium*, we found that sets with smaller distances between primer binding sites on the target genome outperformed those optimized for lower Gini index. Nine out of the ten positive test *Mycobacterium* sets, including the four best sets, had lower Gini indices than the sets for *Wolbachia*. This suggests that after a certain threshold the Gini index becomes secondary to the primer binding site density. Therefore, pre-selecting primers with a low Gini index during **swga filter** and then choosing sets with high binding density in **swga export** allows the optimization of both attributes, and should yield effective primer sets.

3.5. Discussion

Selective whole-genome amplification provides a way to preferentially amplify a target genome from a complex background. However, implementation of the SWGA method has been limited due to the difficulties in designing an effective set of primers. Assembling a primer set where all of the primers are compatible with each other, selective for the target genome, and rare in the background is a problem with many degrees of freedom. The *swga* program addresses this difficulty by automatically identifying and evaluating primer sets by specified criteria, allowing the user to select only those sets most likely to succeed in selective amplification of the target.

We used *swga* to design primer sets for *W. pipientis* and *M. tuberculosis* that selectively amplified each in the presence of their host's genome. These sets had varying binding evenness and selectivity for the target genome, allowing us to compare these attributes to the performance of each set. In addition, we demonstrated potential clinical utility of the *swga* program by amplifying DNA from the *M. tuberculosis* pathogen spiked into human blood. While in these experiments we used target/background pairs with clearly defined genomes, there is no reason the background cannot be a heterogenous mixture of DNA, such as stool or soil. In this case, the background could be approximated by whole-genome shotgun sequencing of the mixture, and subtracting any reads belonging to the target, if present.

Based on these results, it appears that primer binding evenness (as measured by the Gini index), primer set binding selectivity, and the density of binding sites on the target

genome each play an important role in the set's efficacy. In the *W. pipientis* study, we established that the temperature range of 15–45°C for the primers and prioritizing evenness of binding led to more even amplification of the target genome. In addition, the *swga*-designed sets performed better at selectively amplifying the target than the hand-designed set in (Leichty & Brisson, 2014). In fact, the Leichty primer set was not generated by *swga* because the maximum distance between primer sites on the *Wolbachia* genome was greater than the specified cutoff. In *M. tuberculosis*, by starting with a pool of primers that bind relatively evenly to the target, we constrained the range of set binding evenness by removing primers that cluster on repeat regions. After controlling the range of binding evenness at the primer level, the sets with the highest target binding density (i.e. lowest mean distance between binding sites) achieved highest coverage, suggesting that further refinements of the sets for evenness is not necessary. These sets consequently had the lowest ratio of target to background average binding distances. This ratio, as a more complete representation of the set's selectivity than just the binding density on the target, had a strongly inverse correlation with the amount of the genome covered after sequencing (Figure 3-6). Because both attributes are closely related, it is difficult to disentangle the effects of binding density from the effects of a low ratio, and it may be that either or both of these attributes contribute to the success of these primer sets. Furthermore, some sets had relatively similar ratios (e.g. Mtb7 versus Mtb8), but Mtb8 yielded greater genome coverage. This indicates that there are likely other set attributes not considered here that also contribute to set efficacy. To compensate for this,

we suggest selecting five to ten sets with low ratios to test experimentally, and then selecting the best-performing of those sets.

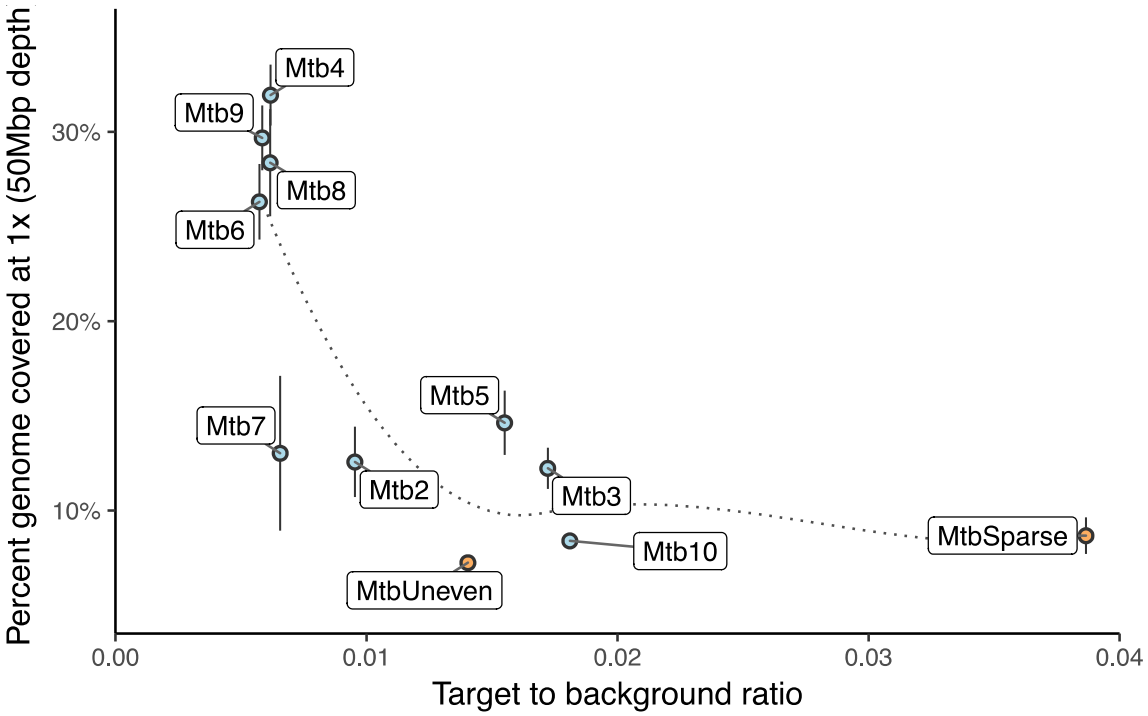


Figure 3-6. Coverage is related to set selectivity.
The percentage of the Mycobacterium genome covered by each set at ×1 coverage after 50 Mbp of sequencing is inversely correlated to the set's target to background binding distance ratio (e.g. selectivity). The smoothed line of best fit (LOESS) is shown by the dotted line. The points and whiskers represent the median and standard deviation of the technical replicates. Positive tests are in blue, while negative controls are in orange. The random hexamers did not have a definable ratio and are not displayed.

The *swga* program does not consider a specific number of primers for each set. Instead, *swga* considers primer sets of different sizes, and reports suggested sets. By exploring a range of sizes, the *swga* program allows the user to find sets with desirable attributes without having to guess what the ideal set size will be in advance.

SWGA is best suited to large scale population genomics studies and may not be cost effective in some smaller studies. Developing the SWGA primer set requires up-front

costs that need to be recovered in later applications for the method to be cost effective. A detailed cost-benefit analysis over multiple applications is presented in Section §3.7.4. SWGA is most useful when large numbers of samples are to be sequenced, when the target genome is rare in the unamplified sample, and when higher sequencing coverage of the target genome is desired.

Our experiments so far suggest a general workflow that can be used to design primer sets for other systems. In particular, we recommend the following guidelines:

- During **swga filter**, set the max_gini parameter as low as possible while still yielding 200 or more primers.
- For **swga find_sets**, set the max_sets to 1–5 million to explore a wide range of set attributes.
- Use **swga export** to export the sets ordered by the distance between binding sites on the target (attribute fg_mean_dist).
- Pick the five to ten sets with lowest fg_mean_dist to test experimentally. Barcode each amplicon separately, then pool and sequence with low depth to assess performance. Once a high-performing set is identified, sequence that amplicon more deeply. This set is now usable in any samples that have similar target/background combinations.

We expect best practices to evolve as SWGA is used more frequently. To facilitate this, we have set up a web page on the project's source repository and a user mailing list. A

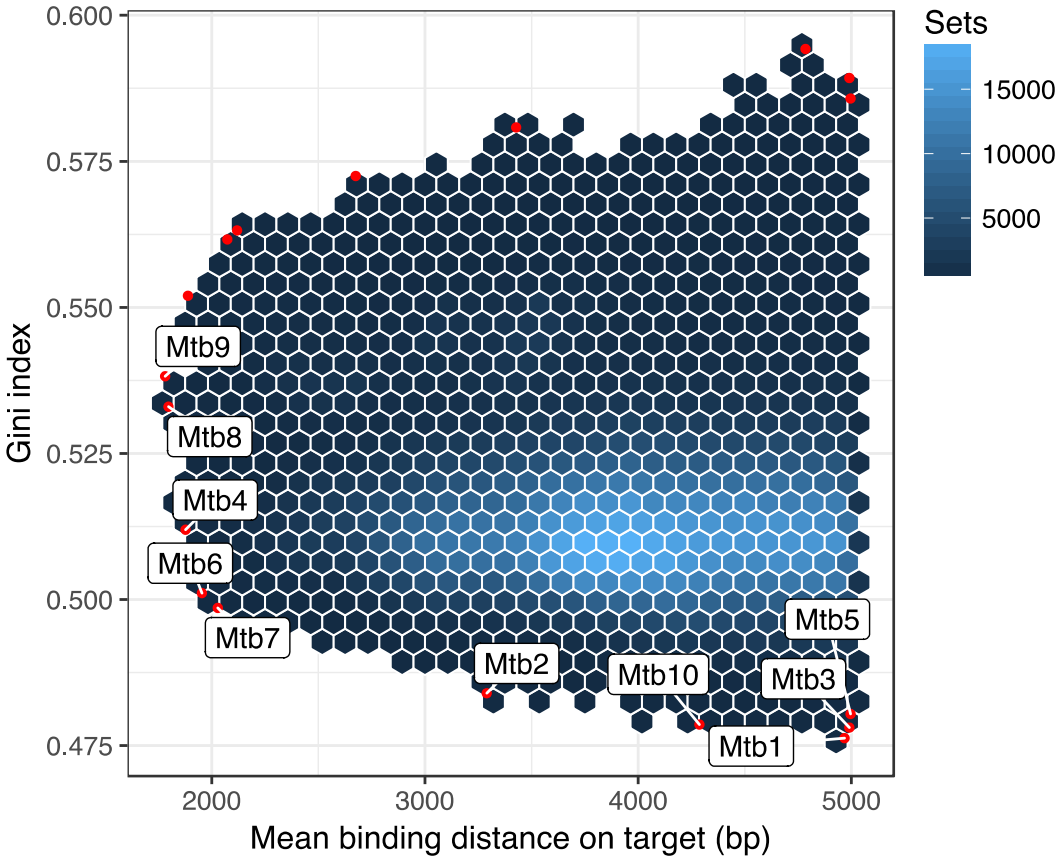
tutorial on the program's operation and more extensive documentation on each parameter and module is available on the web page as well.

3.6. Acknowledgements

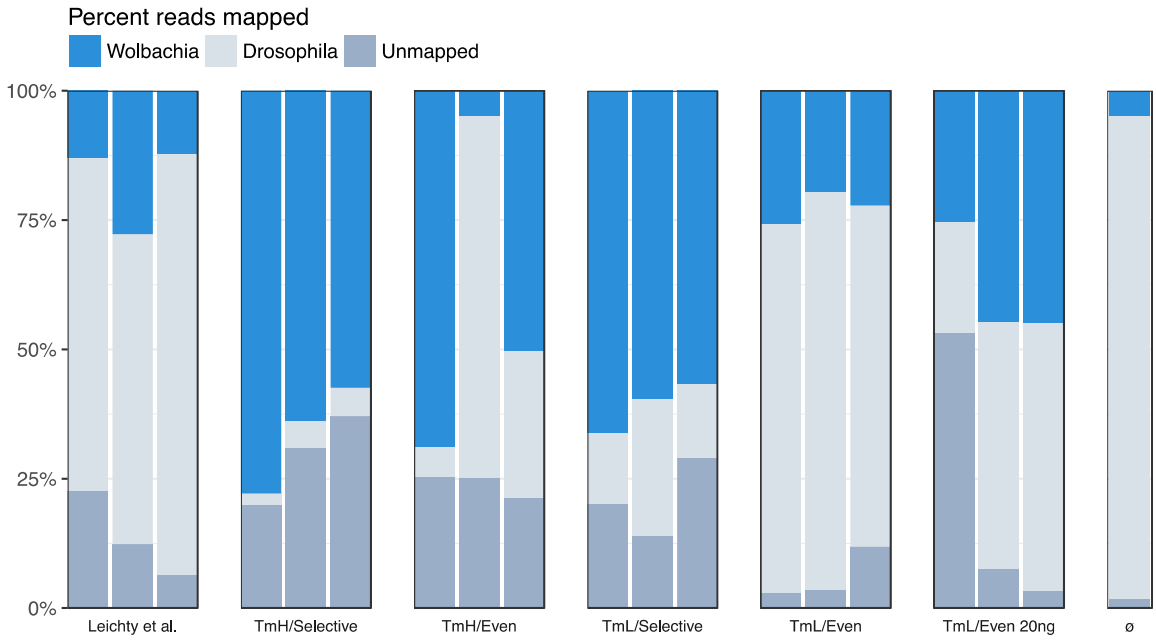
We thank Michael Parisi for his generous donation of *Wolbachia*-infected *Drosophila* strains, as well as Alex Berry and other early *swga* users for their feedback.

3.7. Supplemental Material

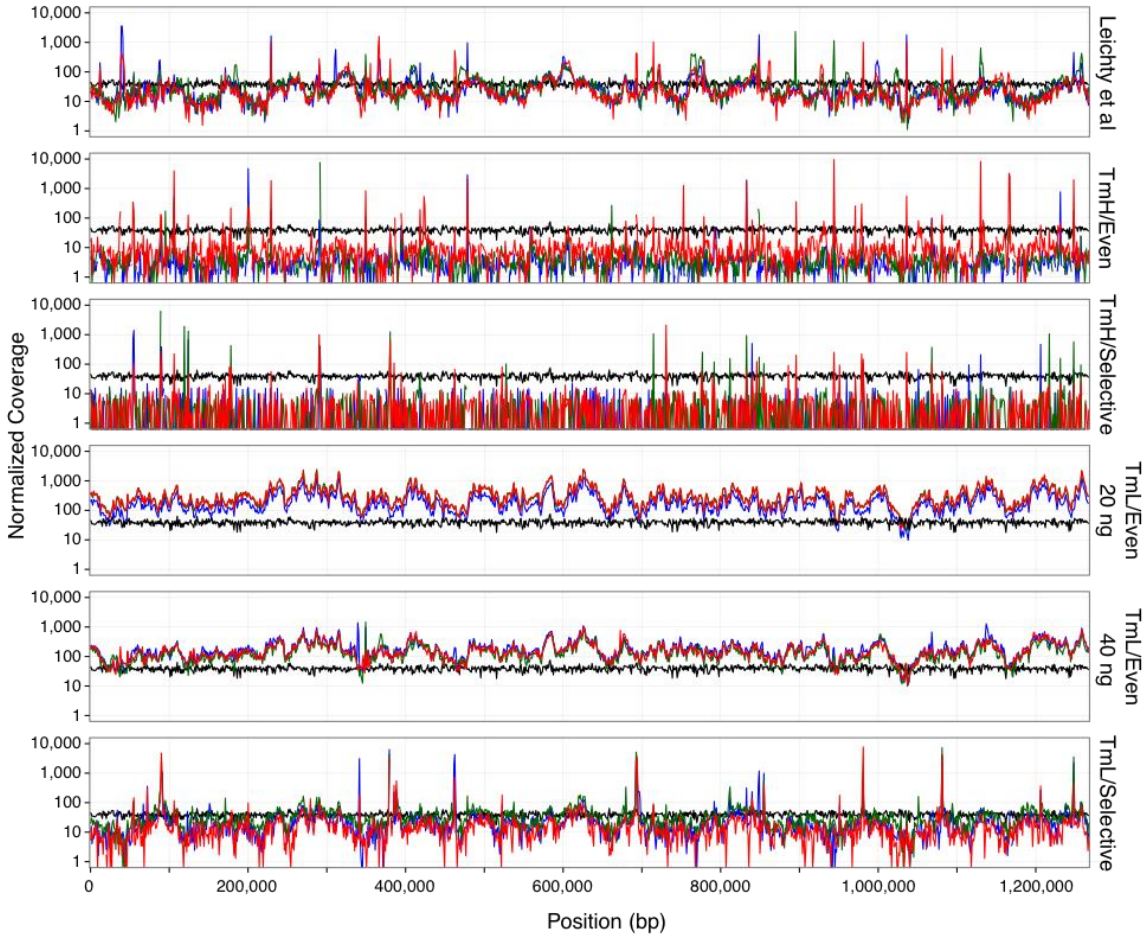
3.7.1. Supplemental Figures



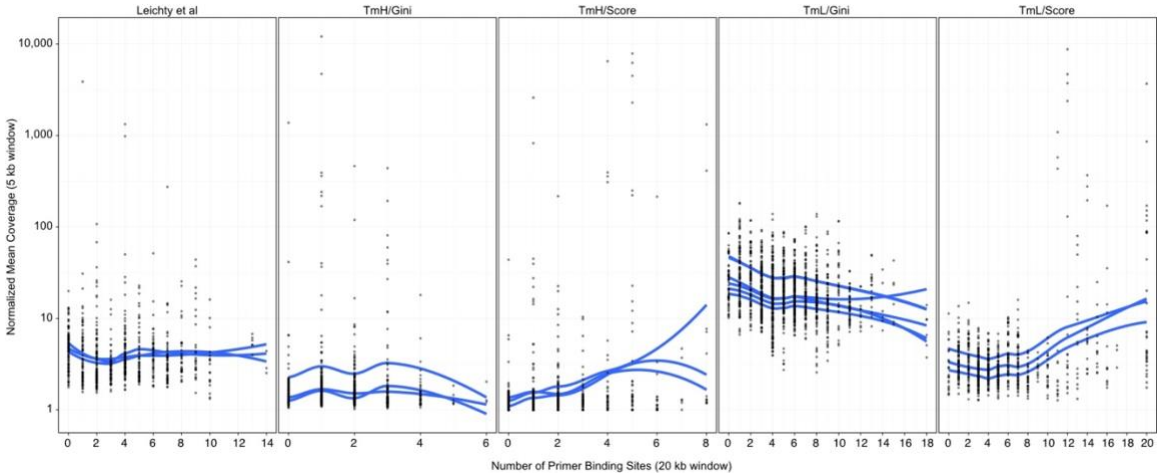
Supp. Figure 3-1.
The sets identified for Mycobacterium tuberculosis with a mean binding distance on the target less than 5kb are shown as a hexplot, where the color intensity represents the number of sets within that range of binding density (x-axis) and evenness (y-axis). We used the 'chull' function in R to detect the inflection points of this distribution. These points (red dots) were the sets that were used for selective amplification. They represent various extreme combinations of primer binding evenness and density. Where there were multiple points close together, we picked one at random.



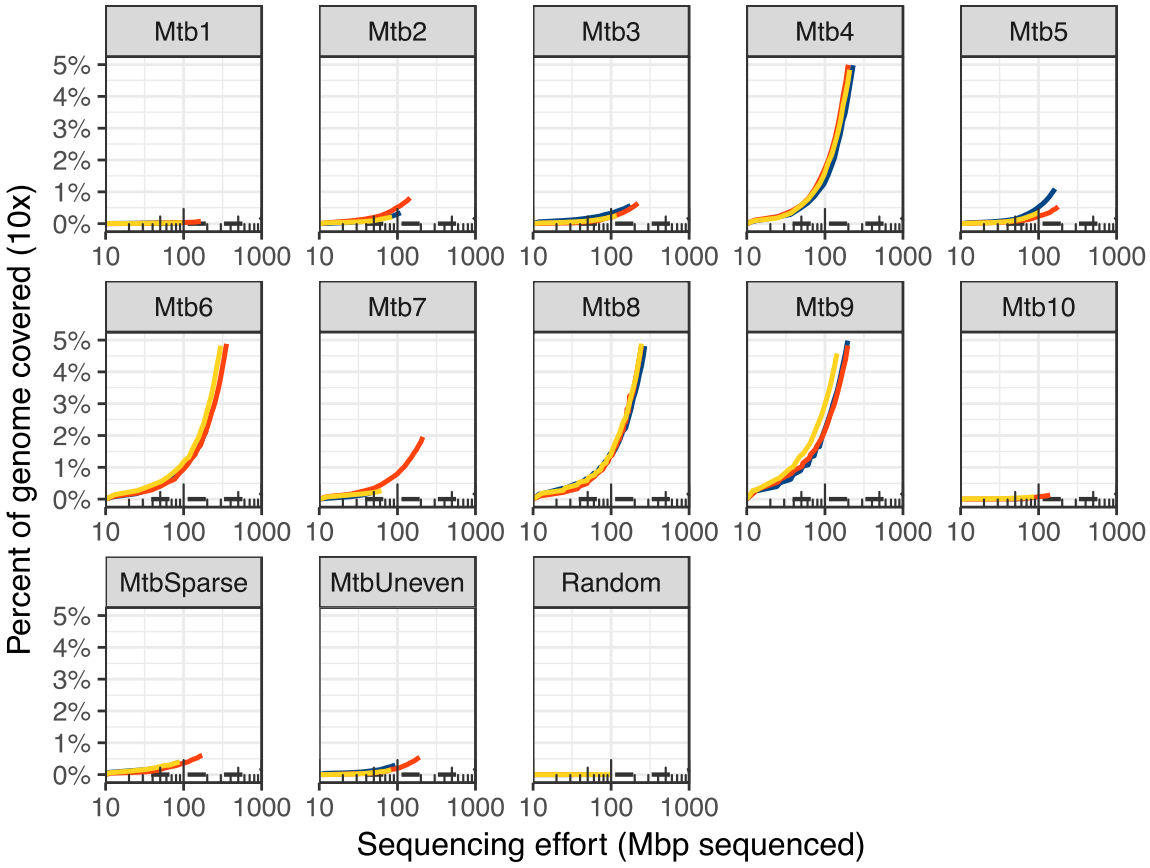
Supp. Figure 3-2.
Selective whole-genome amplification increases the proportion of sequencing reads that map to the target genome. Direct sequencing of total DNA of 10 flies (column labelled ø) resulted in more than 93% of reads mapping to Drosophila while only 4.7% mapped to Wolbachia. By contrast, at least 15% of the reads mapped to Wolbachia after select whole genome amplification with the worst performing primer set. All primer sets identified by the swga program performed substantially better than primer sets chosen manually (Leichty, 2014). Results from each of three replicate SWGA reactions (using 40 ng total DNA per reaction) per primer set are shown. The TmL/Even primer set was also run in triplicate with 20 ng total DNA per reaction. Results were more similar for replicates within a set than between sets, indicating that SWGA performs consistently when using the same primer set.



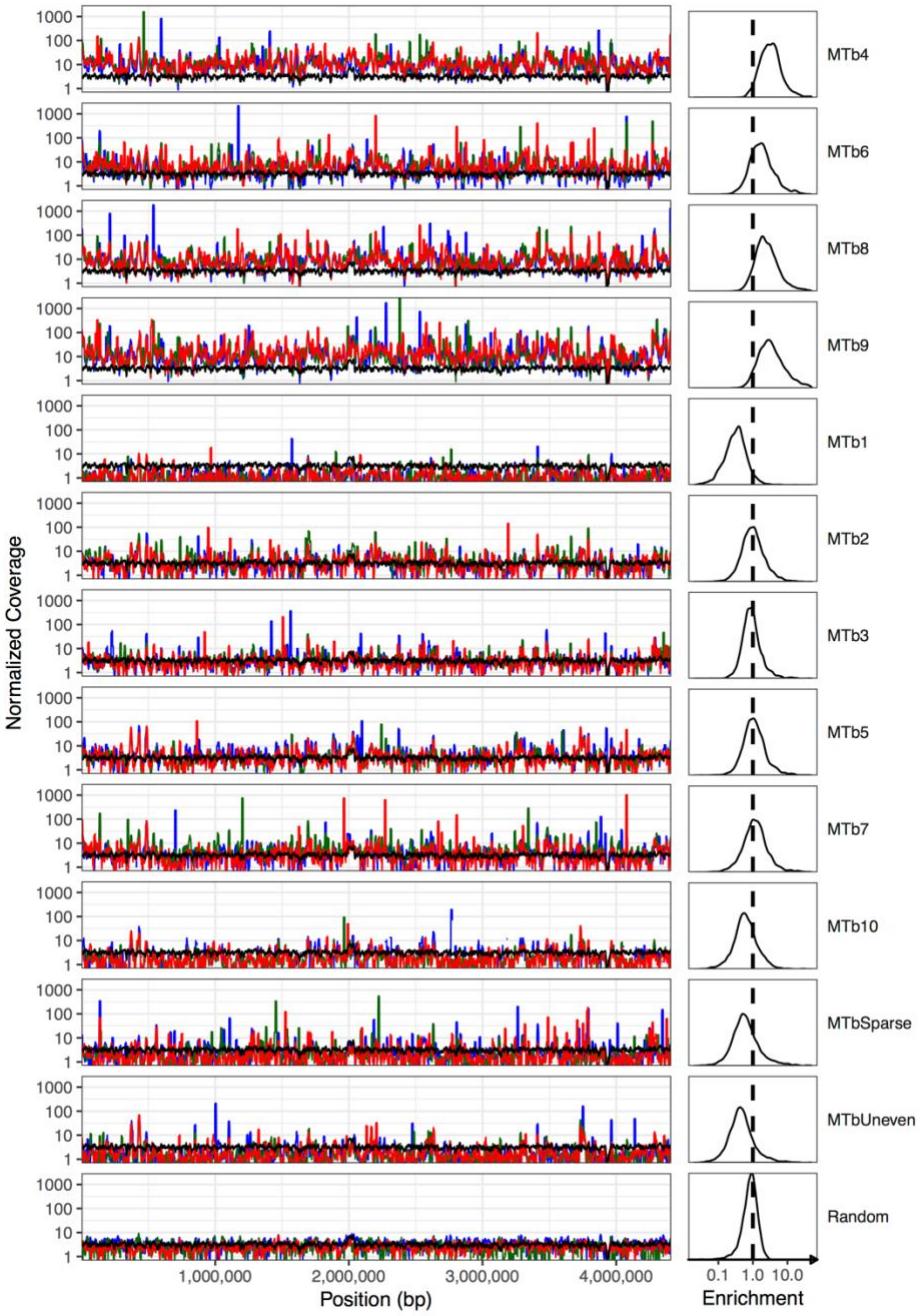
Supp. Figure 3-3.
The depth of sequencing per 1Gb of sequencing effort (1 Gb * coverage depth / total bp sequenced) is shown for all replicates of all primer sets chosen for Wolbachia against Drosophila. Colored lines indicate technical replicates, while the black line indicates the unamplified control. The TmL/Even set was re-sequenced with 20ng of total fly extract DNA in addition to the 40ng used for the other sets. The TmL/Even amplicons consistently achieved higher coverage of the Wolbachia genome than the unamplified control.



Supp. Figure 3-4.
Variation in the local density of primer binding sites does not account for variation in select whole amplification across the Wolbachia genome. The normalized mean sequencing coverage (1 Mb * coverage depth / total bp sequenced) across 5 kb windows after amplification is shown relative to the number of binding sites within a 20 kb window. Smoothed (LOESS) best fit lines (blue lines) are plotted for each technical replicate. Within each primer set, normalized mean coverage did not correlate with the number of binding sites in a 20 kb window. However, coverage did vary between primer sets and was generally higher for primer sets with a higher number of primer binding sites (Table 1).



Supp. Figure 3-5.
10x coverage of the Mycobacterium genome was limited without additional sequencing effort. Colored lines indicate technical replicates, while black dashed lines indicate the unamplified control.



Supp. Figure 3-6.
The depth of sequencing per 1Gb of sequencing effort (1 Gb * coverage depth / total bp sequenced) is shown for all replicates of all primer sets chosen for Mycobacterium against the human genome. The first four sets yielded higher coverage of the Mtb genome than the unamplified control, as indicated by each the colored line for each replicate being higher than the unamplified control (black line). For each set, a density histogram of the ratio between the average set coverage and the unamplified control is shown on the right. Peaks to the right of the dashed line (indicating a ratio greater than 1) represent overall enrichment.

3.7.2. Supplemental Tables

Supp. Table 3-1. Wolbachia primer sets

ID	Ratio	Size	Gini	Max target dist	Mean target dist.
TmL/Selective	0.0544	9	0.65	3.11E+04	5.33E+03
TmL/Even	0.1050	7	0.54	3.47E+04	6.85E+03
TmH/Even	0.0075	2	0.54	1.13E+05	1.31E+04
TmH/Selective	0.0005	2	0.66	1.28E+05	1.21E+04
Leichty and Brisson 2014	0.0163	2	0.71	1.13E+05	5.31E+03

ID	Target dist (sd)	Mean bg. dist	Primers
TmL/Selective	7.00E+03	9.78E+04	AACATAGATC,AAGAGATACC, AATGTTCGTA, ACGTGTTAG, AGAAATTTACTA, ATCTAGAGAT, ATCTATGTTAAG, TACGTCATAC, TATGCAAGAA
TmL/Even	7.00E+03	6.53E+04	AATGTTCGTA, ATAAGCTGAA, TAAAGACATA, TACGTCATAC, TATGCAAGAA, TGAGATACC, TTTCTGGATC
TmH/Even	1.55E+04	1.73E+06	CACTGGAATCC, CGCTACTTGTTA
TmH/Selective	1.80E+04	2.43E+07	CTACGTGTTAGC, CTACTTGTTAGC
Leichty and Brisson 2014	1.01E+04	3.25E+05	ATCCAAGTAG, CGGTATCTC

Supp. Table 3-2. Mycobacterium primer sets.

Name	ID	Ratio	Size	Gini	Max target dist	Mean target dist.
Mtb1	1130935	0.0171	7	0.476	3.27E+04	4.97E+03
Mtb2	1236643	0.0095	7	0.484	2.75E+04	3.29E+03
Mtb3	1482488	0.0172	7	0.478	3.77E+04	4.99E+03
Mtb4	1558358	0.0062	7	0.512	1.49E+04	1.88E+03
Mtb5	1951361	0.0155	6	0.480	3.44E+04	5.00E+03
Mtb6	4690256	0.0057	7	0.501	1.50E+04	1.95E+03
Mtb7	4715948	0.0066	7	0.499	1.39E+04	2.03E+03
Mtb8	5192056	0.0062	7	0.533	1.87E+04	1.80E+03
Mtb9	5194179	0.0058	7	0.538	1.87E+04	1.78E+03
Mtb10	5699436	0.0181	7	0.479	3.12E+04	4.29E+03
MtbUneven	70336	0.0140	2	0.623	9.21E+04	1.14E+04
MtbSparse	146196	0.0387	3	0.505	1.00E+05	2.60E+04

Name	Target Dist (sd)	Mean bg. dist	Primers
Mtb1	4.56E+03	2.90E+05	ACGATCA*A*C,CCGATAT*G*G,CGCGAA*T*A,CGCGAT*T*A,CGTCGT*A*G,CGTCGT*A*T,TAGTCGA*T*G
Mtb2	3.12E+03	3.45E+05	ATTCGT*C*G,CGCGAA*T*A,CGGTAT*C*G,CGTCGT*A*A,CGTCGT*A*T,GATTGTC*G*A,TACGAA*C*G
Mtb3	4.69E+03	2.90E+05	ATCGGAT*T*C,CGATAC*G*T,CGCGAT*A*A,CTACGA*C*G,CTCGATA*C*C,GATCGAC*T*C,TCGATCA*A*C
Mtb4	1.92E+03	3.04E+05	ATCGACA*A*C,CGAATC*C*G,CGTTAC*G*G,CTACGA*C*G,GACGAT*C*G,GATCGAC*T*C,TCGACG*A*A
Mtb5	4.71E+03	3.22E+05	ATATCGG*T*G,CCGAAT*C*G,CGGTTA*C*G,GACGACT*A*C,GATGATC*G*A,TGGATAT*C*G
Mtb6	1.95E+03	3.41E+05	CCGAAT*C*G,CGCGAA*T*A,CGCTAT*C*G,CGGTAT*C*G,CGGTTA*C*G,CGTCTA*C*G,TCGACG*A*A
Mtb7	1.99E+03	3.09E+05	ATCGACA*A*G,CCGAAT*C*G,CGCGAA*T*A,CGCTAT*C*G,CGGTAT*C*G,CGTCTA*C*G,TCGACG*A*A
Mtb8	1.96E+03	2.92E+05	ACGATCA*A*C,CGAATC*C*G,CGACGA*A*A,CGACGA*T*A,CGATAC*C*G,TACGAC*G*A,TCGACG*A*A
Mtb9	1.96E+03	3.05E+05	CGAATC*C*G,CGACGA*A*A,CGACGA*T*A,CGATAA*C*G,CGATAC*C*G,TACGAC*G*A,TCGACG*A*A
Mtb10	4.09E+03	2.37E+05	CGAAAC*G*A,CGATATC*C*A,CGATTC*G*G,CGGTATT*G*A,CGTAGT*C*G,CGTTAC*G*T,TGTAGTC*G*A
MtbUneven	1.45E+04	8.10E+05	CGATATT*G*C,CGTTAC*C*G
MtbSparse	2.48E+04	6.71E+05	CGTTCG*T*A,TGATATC*G*C,TTCGACT*A*C

3.7.3. Supplemental Data

3.7.3.1. Swga settings for Wolbachia pipientis against Drosophila melanogaster

```
[count]
min_size = 5
max_size = 12
min_fg_bind = 20
max_bg_bind = 10000
max_dimer_bp = 4
exclude_threshold = 1

[summary]
bg_length = 145523498
fg_length = 1285894

[filter]
max_primers = 200
bg_length = 145523498
fg_length = 1285894
min_fg_bind = 41
max_bg_bind = 474
min_tm = 15
max_tm = 45

[find_sets]
min_size = 2
max_size = 12
max_dimer_bp = 4
min_bg_bind_dist = 30000
bg_genome_len = 145523498
max_fg_bind_dist = 36000
max_sets = 1
workers = 3

[score]
bg_genome_len = 145523498
score_expression = (fg_dist_mean * fg_dist_gini) / (bg_dist_mean)

[export]
window_size = 10000
```

3.7.3.2. Swga settings for Mycobacterium tuberculosis against Homo sapiens.

```
[count]
min_size = 5
max_size = 12
min_fg_bind = 44
max_bg_bind = 21770
max_dimer_bp = 3
exclude_threshold = 1

[export]
window_size = 10000

[filter]
max_primers = 200
min_fg_bind = 50
max_bg_bind = 21770
min_tm = 15
max_tm = 45
max_gini = 0.6

[find_sets]
min_size = 2
max_size = 7
max_dimer_bp = 3
min_bg_bind_dist = 60000
max_fg_bind_dist = 100000
max_sets = -1
workers = 4
score_expression = (fg_dist_mean * fg_dist_gini) / (bg_dist_mean)

[score]
score_expression = (fg_dist_mean * fg_dist_gini) / (bg_dist_mean)
```

3.7.3.3. Code used in this project

For code used in analysis and figures, please refer to
https://github.com/eclarke/swga_paper

3.7.4. Cost analysis for SWGA

The SWGA method improves the efficiency of genomic sequencing when targeting the genome of a specific organism in the presence of contaminating DNA. However, implementing SWGA itself incurs costs, and so not all applications will benefit from its use. In particular, SWGA is generally more cost effective studies that require sequencing a large fraction of the target genome from multiple samples. For experiments necessitating new primer design, as described in this study, either the depth of sequencing coverage or number of samples must be large enough to offset initial cost of purchasing

and testing primer sets. For these reasons, SWGA, using newly designed primer sets, is not necessarily cost effective for applications with very small target genomes, low coverage depth , or small numbers of samples.

Here we present a cost analysis, including the cost of de novo primer set design and testing. To highlight the areas where SWGA is most cost effective, we have performed this analysis across range of sequencing depths for three target organisms: *Mycobacterium tuberculosis* in humans, *Plasmodium vivax* in humans, and *Wolbachia pipientis* in fruit flies. The cost of implementing and performing SWGA is compared directly to achieving the same depth and coverage of sequencing by direct sequencing of the sample without SWGA amplification.

For each target, we compare the cost of SWGA sequencing for a given number of samples, including primer set design and testing, to that of sequencing directly from the starting DNA mixture. We calculate the cost of sequencing samples at 1X and 10X coverage and with the goal of covering 50%, 75%, and 80% of the target genome, thus illustrating how the cost effectiveness of SWGA varies with depth and breadth of sequencing.

The following is a breakdown of the costs associated with sequencing and the SWGA process. All costs are in US dollars, and reagent prices are accurate as of the time of writing (Feb 2017).

Supp. Table 3-3. One-time costs associated with SWGA.

Step	Unit	Cost
Primer design	Computation time*	$5.88
Primer testing	Miseq (reagents + sequencing)	$1,500.00

Computation time measured as the cost of running swga on the Penn High Performance Computing cluster on one node for one week at $0.035/hr/node.

Supp. Table 3-4. Per-sample costs associated with SWGA.

Step	Unit	Cost/sample
Primer Set (average for 10 primers)	Primer oligos	$0.02
SWGA Reaction	Phi29 + Buffer	$5.70
SWGA Reaction	DNTPs	$0.41
Sequencing (Nextera - Hiseq)	Kit reagents	$65.00
Cost per base on an Illumina Hiseq using Nextera: $0.000000094.		

Total cost is $71.13/sample. Costs per sample were estimated from the cost of and number of uses yielded by the reagent (for specific reagent catalog numbers, see the main text).

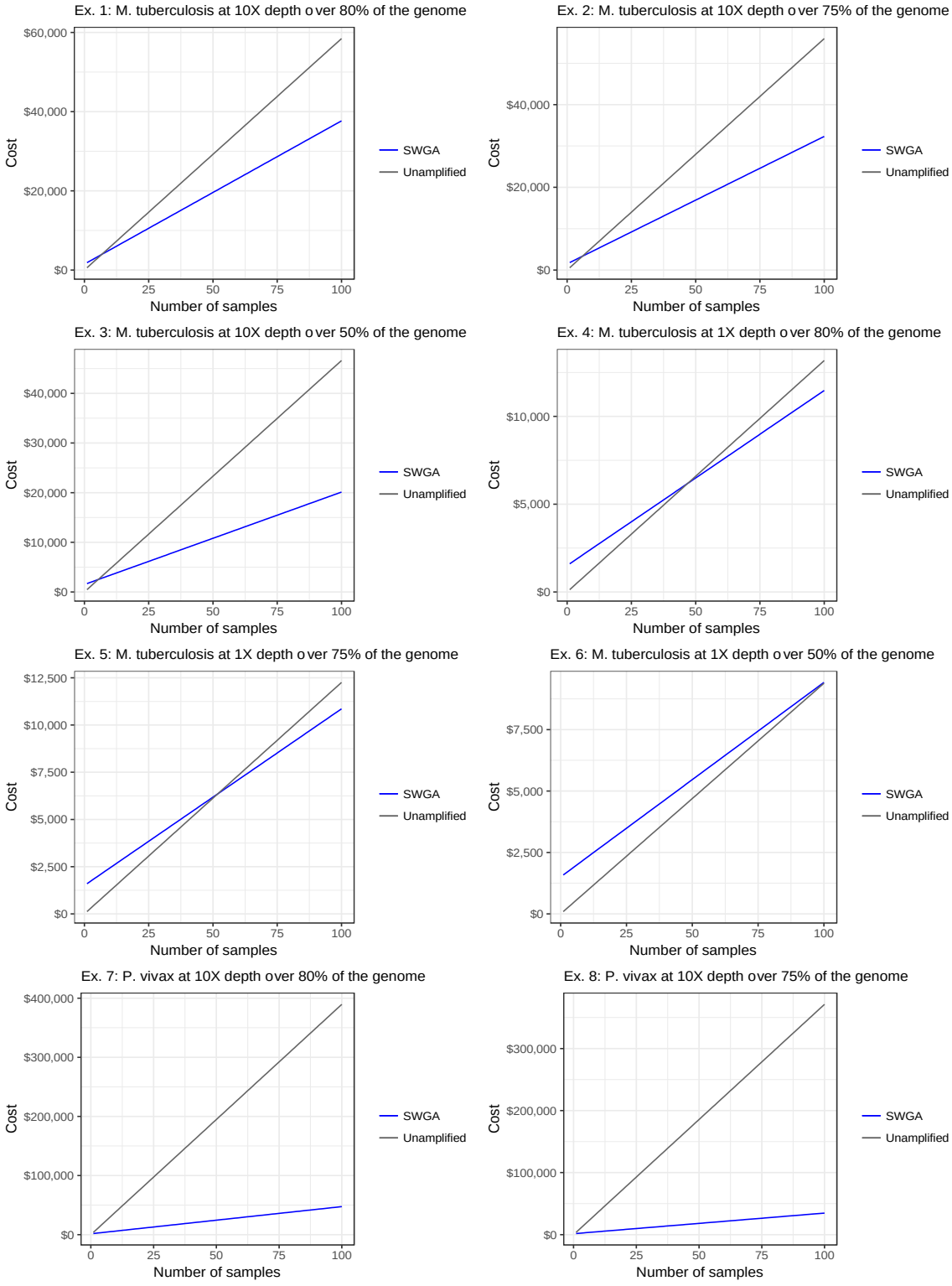
To calculate the cost of sequencing some number of samples n with SWGA, we added the fixed costs (C_f) and multiplied the variable costs C_v by the number of samples. We then empirically determined the number of bases required to achieve the required depth of coverage (N_{bp}, see Table 2 at end of document). This was multiplied by the above cost per base (C_{bp})and the number of samples:

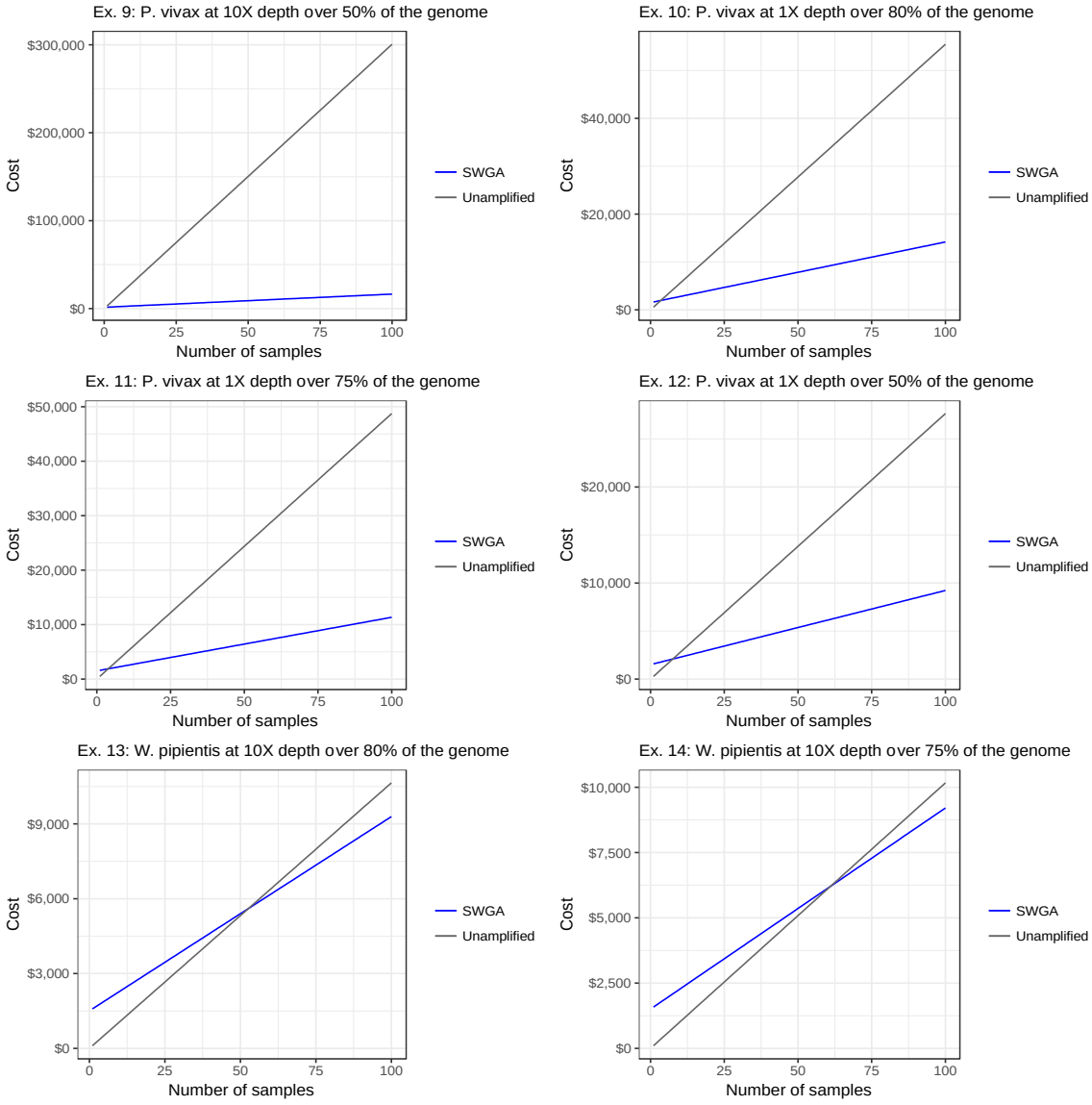
$$\mathrm{Cost}_{\mathrm{SWGA}}(n) = C_f + (C_v \times n) + (N_{bp} \times C_{bp} \times n)$$

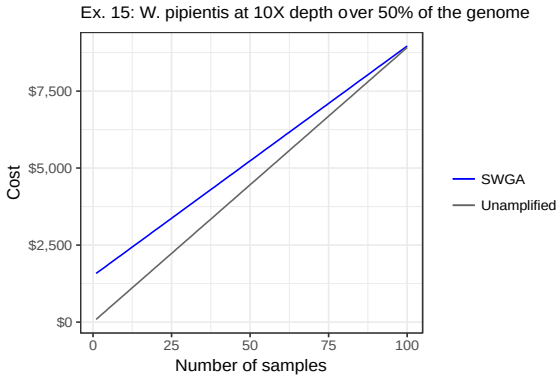
The cost of directly sequencing a sample (without SWGA) is the cost of the sequencing reagents (C_r) and the sequencing itself:

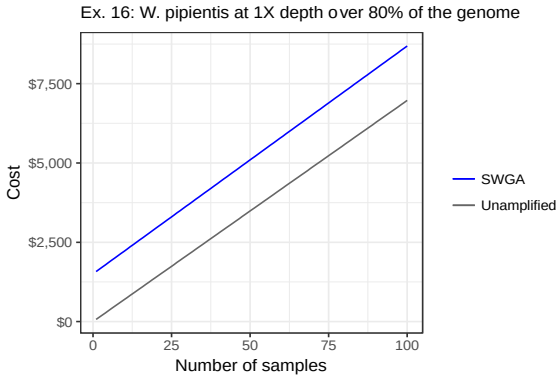
$$\mathrm{Cost}_{\mathrm{Unamplified}}(n) = (C_r \times n) + (N_{bp} \times C_{bp} \times n)$$

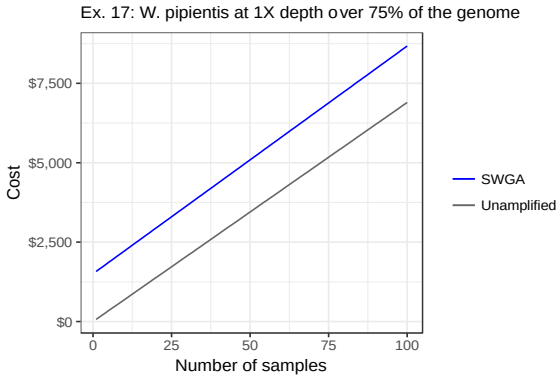
In each plot, the intersection point between the two lines (SWGA vs unamplified) represents the point at which the SWGA method becomes more cost efficient. This is heavily influenced by the relative efficiency gains yielded by SWGA, which itself is a function of the target and host genomes. For some applications, like sequencing *W. pipientis* at low coverage depth and breadth, SWGA is not cheaper than conventional sequencing.

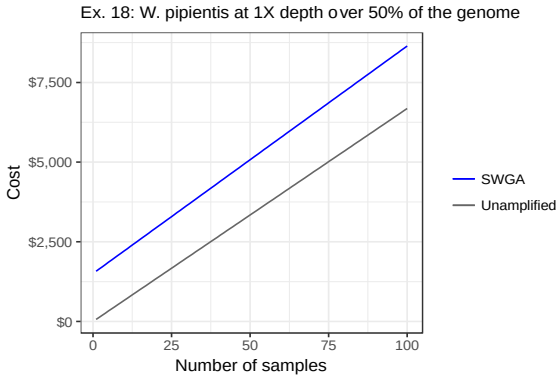

Ex. 1: M. tuberculosis at 10X depth over 80% of the genome
Ex. 2: M. tuberculosis at 10X depth over 75% of the genome
Ex. 3: M. tuberculosis at 10X depth over 50% of the genome
Ex. 4: M. tuberculosis at 1X depth over 80% of the genome
Ex. 5: M. tuberculosis at 1X depth over 75% of the genome
Ex. 6: M. tuberculosis at 1X depth over 50% of the genome
Ex. 7: P. vivax at 10X depth over 80% of the genome
Ex. 8: P. vivax at 10X depth over 75% of the genome
Cost
Number of samples
SWGA
Unamplified


Ex. 9: P. vivax at 10X depth over 50% of the genome
$300,000
$200,000
$100,000
$0
Cost
0
25
50
75
100
Number of samples
SWGA
Unamplified
Ex. 10: P. vivax at 1X depth over 80% of the genome
$40,000
$20,000
$0
Cost
0
25
50
75
100
Number of samples
SWGA
Unamplified
Ex. 11: P. vivax at 1X depth over 75% of the genome
$50,000
$40,000
$30,000
$20,000
$10,000
$0
Cost
0
25
50
75
100
Number of samples
SWGA
Unamplified
Ex. 12: P. vivax at 1X depth over 50% of the genome
$20,000
$10,000
$0
Cost
0
25
50
75
100
Number of samples
SWGA
Unamplified
Ex. 13: W. pipientis at 10X depth over 80% of the genome
$9,000
$6,000
$3,000
$0
Cost
0
25
50
75
100
Number of samples
SWGA
Unamplified
Ex. 14: W. pipientis at 10X depth over 75% of the genome
$10,000
$7,500
$5,000
$2,500
$0
Cost
0
25
50
75
100
Number of samples
SWGA
Unamplified


Ex. 15: W. pipientis at 10X depth over 50% of the genome
Cost
$7,500
$5,000
$2,500
$0
0
25
50
75
100
Number of samples
SWGA
Unamplified


Ex. 16: W. pipientis at 1X depth over 80% of the genome
Cost
$7,500
$5,000
$2,500
$0
0
25
50
75
100
Number of samples
SWGA
Unamplified


Ex. 17: W. pipientis at 1X depth over 75% of the genome
Cost
$7,500
$5,000
$2,500
$0
0
25
50
75
100
Number of samples
SWGA
Unamplified


Ex. 18: W. pipientis at 1X depth over 50% of the genome
Cost
$7,500
$5,000
$2,500
$0
0
25
50
75
100
Number of samples
SWGA
Unamplified

Supp. Table 3-5. Sequencing required to achieve desired coverage.

Example	Organism	Depth	Coverage	SWGA (bp)	Unamplified (bp)
1	*M. tuberculosis*	10	80	3,091,559,998	5,527,190,382
2	*M. tuberculosis*	10	75	2,520,410,974	5,262,496,048
3	*M. tuberculosis*	10	50	1,223,345,194	4,267,942,796
4	*M. tuberculosis*	1	80	303,869,124	710,912,333
5	*M. tuberculosis*	1	75	238,173,500	612,325,622
6	*M. tuberculosis*	1	50	85,299,735	307,065,400
7	*P. vivax*	10	80	4,129,854,209	40,750,765,958
8	*P. vivax*	10	75	2,780,236,529	38,803,668,507
9	*P. vivax*	10	50	837,785,449	31,291,283,092
10	*P. vivax*	1	80	590,545,217	5,210,647,250
11	*P. vivax*	1	75	289,985,276	4,493,713,988
12	*P. vivax*	1	50	64,803,452	2,249,610,375
13	*W. pipientis*	10	80	71,621,179	441,055,975
14	*W. pipientis*	10	75	62,712,293	390,478,303
15	*W. pipientis*	10	50	36,527,254	256,900,749
16	*W. pipientis*	1	80	7,234,721	50,185,804
17	*W. pipientis*	1	75	5,946,277	42,309,224
18	*W. pipientis*	1	50	2,451,818	19,419,362

3.7.4.1. Conclusions

As shown in Examples 1-3 and 7-12, the cost effectiveness of SWGA is more significant when pursuing higher coverage depth across the target genome. In *M. tuberculosis*, SWGA becomes more cost effective when considering sequencing more than 8 samples. In *P. vivax* SWGA is nearly always more cost effective, representative of the variable effect the target and host genomes can play on the final efficiency gains.

For low-depth sequencing, and for small genomes such as *W. pipientis*, the benefit is only realized at higher numbers of samples. Low-depth sequencing of *M. tuberculosis* is only more cost-effective with SWGA when more than 65 samples are being sequenced (Ex 4-6). Again indicative of the variable influence of the target genome size, however, low-depth coverage of *P. vivax* is more cost-effective with SWGA by only 5 or so samples. As a contrast, it is unlikely that SWGA would be cost-effective at all for low-depth sequencing of *W. pipientis* given that the lines in Examples 16-18 will only converge after a very large (>10,000) number of samples. Importantly, these graphs include the cost of designing and testing new SWGA primers. In situations where primer sets already exist, SWGA would become cost-effective at even lower numbers of samples.

These graphs show that for a variety of target/host genomes, SWGA's benefits are greatest when 1) large numbers of samples are studied, 2) when high depths of coverage are needed, and 3) a large proportion of the genome should be covered.

Chapter 4. T cell dynamics and response of the microbiota after SCID-X1 gene therapy

This chapter is based upon work by Erik Clarke[1], Frances Male, Andrew Connell[1], Nadia Kadry[1], Arwa Abbas[1], Young Hwang[1], John Everett[1], Casey Hofstaedter, Judith Kelsen, Marina Cavazzana, Emmanuelle Six, Alain Fischer, Luigi Notarangelo, Salima Hacein-Bey Abina, Don Kohn, David Williams, Sung-Yun Pai, and Frederic D. Bushman[1].

1. Dept. of Microbiology, University of Pennsylvania School of Medicine

4.1. Abstract

Mutation of the IL2RG gene results in a severe combined immune deficiency (SCID-X1) that has been treated successfully with hematopoietic stem cell (HSC) gene therapy. SCID-X1 gene therapy results in reconstitution of the previously lacking T-cell compartment, allowing analysis of the roles of T-cells in humans by comparing before and after gene correction. Here we interrogate T-cell reconstitution using four forms of high content analysis. 1) Estimation of the numbers of transduced progenitor cells by monitoring unique positions of integration of the therapeutic gene transfer vector. 2) Estimation of T-cell population structure by sequencing of the TCR-beta VDJ-recombination products. 3) Metagenomic analysis of microbial populations in oropharyngyl, nasopharyngal and gut samples. 4) Metagenomic analysis of viral populations in gut samples. Comparison of progenitor and T-cell populations allows estimation of a minimum number of cell divisions needed to generate the observed populations. Analysis of microbial populations shows the effects of immune reconstitution, including normalization of gut microbiota and clearance of viral infections. Metagenomic analysis revealed a notable enrichment of genes for antibiotic

resistance in gene-corrected subjects relative to healthy controls. These data highlight the novel analytical avenues made possible by successful SCID-X1 gene therapy.

4.2. Introduction

Several primary immunodeficiencies have now been treated successfully by gene-correction of hematopoetic stem cells (HSC) with integrating vectors (Aiuti et al., 2013; Aiuti et al., 2007; Aiuti et al., 2002; Biffi et al., 2013; Marina Cavazzana-Calvo, Andre-Schmutz, & Fischer, 2013; Gaspar et al., 2011; Gaspar et al., 2004; Hacein-Bey-Abina et al., 2010; Hacein-Bey-Abina et al., 2002). This work has benefited many patients, and in addition provides a unique window on immune mechanisms. In SCID-X1, the first primary immunodeficiency treated successfully by gene transfer, patients harbor mutations in the IL2RG gene, which encodes the common gamma chain, a component of several cytokine receptors important in T and NK-cell growth and development(Kovanen & Leonard, 2004; Noguchi et al., 1993; Puck et al., 1993). Patients typically lack these cells before correction(Kennedy et al., 2000; Lodolce et al., 1998; Puel, Ziegler, Buckley, & Leonard, 1998), but afterwards show robust T reconstitution and transient NK-cell reconstitution, accompanied by restoration of considerable immune function (Gaspar et al., 2011; Gaspar et al., 2004; Hacein-Bey-Abina et al., 2010; Hacein-Bey-Abina et al., 2002). Gene correction thus provides a unique opportunity to study the onset of T cell function in previously deficient human subjects.

In the first gene therapy trial to treat SCID-X1, early designs of gammaretroviral vectors were used (Gaspar et al., 2011; Gaspar et al., 2004; Hacein-Bey-Abina et al., 2010;

Hacein-Bey-Abina et al., 2002), which were the only vector type available at the time. These vectors contain strong enhancers derived from the starting retroviral backbone. The enhancers supported efficient expression of the corrective IL2RG gene and allowed successful gene correction. However, subsequent experience implicated these vectors in insertional mutagenesis, in which vector signals activated transcription of host proto-oncogenes, in some cases associated with severe adverse events(Hacein-Bey-Abina, von Kalle, Schmidt, Le Deist, et al., 2003; Howe et al., 2008).

A second trial was carried out to treat SCID-X1 using an improved vector in which the strong enhancer sequences were deleted (Hacein-Bey-Abina et al., 2014), and a more specific promoter was used to express the therapeutic IL2RG gene copy. T-cell numbers after correction were indistinguishable in the first and second trials. So far, no severe adverse events have been linked to insertional activation in the second trial.

In this study, we used multiple high throughput sequence-based methods to analyze samples from the SCID-X1 trials, with the goal of probing immune mechanisms and the resulting effects on microbial communities. To assess the number and distributions of gene-corrected precursor cells producing T-cells, deep sequencing of sites of vector integration was used (C. Berry, Hannenhalli, Leipzig, & Bushman, 2006; C. C. Berry et al., 2012; C. C. Berry et al., 2017; Hacein-Bey-Abina et al., 2014; Hacein-Bey-Abina, von Kalle, Schmidt, Le Deist, et al., 2003; Hacein-Bey-Abina, Von Kalle, Schmidt, McCormack, et al., 2003; Howe et al., 2008; Mitchell et al., 2004; Schroder et al., 2002; Sherman et al., 2017; G. P. Wang et al., 2010; G. P. Wang et al., 2007), where each

unique integration site marked a distinct T-cell progenitor. T-cell development could be followed at a later step by using DNA sequencing to track rearrangements of gene segments encoding the T-cell receptor-beta CDR3 region (Boyd et al., 2009; Campregher, Srivastava, Deeg, Robins, & Warren, 2010; H. S. Robins et al., 2009; H. S. Robins et al., 2010; Weinstein, Jiang, White, Fisher, & Quake, 2009). Immune cells contribute to control of the resident microbiota, so the consequences of T-cell reconstitution were assessed by deep sequencing of oral, fecal and nares samples to characterize the full microbiota using shotgun metagenomics. In a separate analysis, samples were enriched biochemically for viral particles, RNA and DNA extracted, and then the viral content monitored in fecal samples.

The data support a wealth of new inferences on T-cell growth and immune activity after reconstitution. For example, a minimum estimate for the numbers of cell divisions between progenitor cells and mature T-cells was developed by comparing population sizes from integration site and TCR sequence data. In the microbiome data, normalization of microbial communities was documented following successful treatment in several subjects. Viral infections, several not detected clinically, could be shown to be widespread but often cleared with immune reconstitution. TCR diversity could be compared for selected samples from the SCID1 and SCID2 trails, providing information on the durability of reconstitution and effects of adverse events on TCR diversity. Thus these data begin to outline the utility of "multi-omics" analysis of gene correction in primary immunodeficiency.

4.3. Results

4.3.1. Experimental strategy

Our comparative analysis of gene-corrected progenitors and daughter T cells focused on four patients from the SCID-2 trail (Hacein-Bey-Abina et al., 2014) and three patients from the SCID-1 trial (Hacein-Bey-Abina et al., 2010) for whom samples were available Figure 4-1. Patient characteristics are summarized in Supp. Table 4-1. Additional subjects were studied for which we did not have T cell samples, but did have peripheral blood mononuclear cell samples (PBMC, Supp. Figure 2-1). Adverse events took place in the SCID-1 trial involving subjects F107 and F110 at times 68 months and 33 months after infusion (Hacein-Bey-Abina et al., 2010; Hacein-Bey-Abina et al., 2014; G. P. Wang et al., 2010). The time points analyzed here are well after these adverse events, allowing assessment of the effects of leukemia and chemotherapy on progenitor cell and T cell populations. One subject in the SCID2 trial, B205, was transplanted twice without achieving clinical reconstitution. Samples from the two unsuccessful treatments are designated B205 and B205b.

Vector integration sites and the T cell repertoire were monitored through regular per-protocol blood draws, followed by isolation of CD3+ populations. Sorted cells were subject to DNA extraction and then amplification of either integration sites (mostly previously reported in (Hacein-Bey-Abina et al., 2010; Hacein-Bey-Abina et al., 2014; G. P. Wang et al., 2010); Supplementary Table X) or mature TCR beta loci (Supp. Table

4-3; data new here), followed by sequencing. Vector integration sites mark progenitor cells capable of delivering mature T-cells to the periphery. Rearranged TCR beta loci mark mature T-cells present in blood. Time points from the integration site analysis were chosen to match those used in the TCR analysis—characterization of additional time points for integration site distributions in SCID-X1 gene therapy can be found in (M. Cavazzana-Calvo et al., 2000; Hacein-Bey-Abina et al., 2010; Hacein-Bey-Abina et al., 2014; Howe et al., 2008; Thrasher et al., 2006; G. P. Wang et al., 2010; G. P. Wang et al., 2008).

Microbiome samples were available for six SCID-2 patients (B201, B203, B204, B205, B207, and F201). Oral, nasal, and gut microbiota were sampled via collection of oropharyngeal swabs, nasopharyngeal swabs, and stool samples. Sampling times ranged from 4 to 181 months post infusion of corrected cells. Sample acquisition was at times limited by clinical and practical considerations.

As controls, we analyzed TCR-beta sequences from CD3+ cells from five healthy children and three healthy adults. Healthy subject demographics are in Supp. Table 4-1. Cross-sectional microbiome samples were collected from the same subjects and analyzed as for the SCID gene-corrected samples.

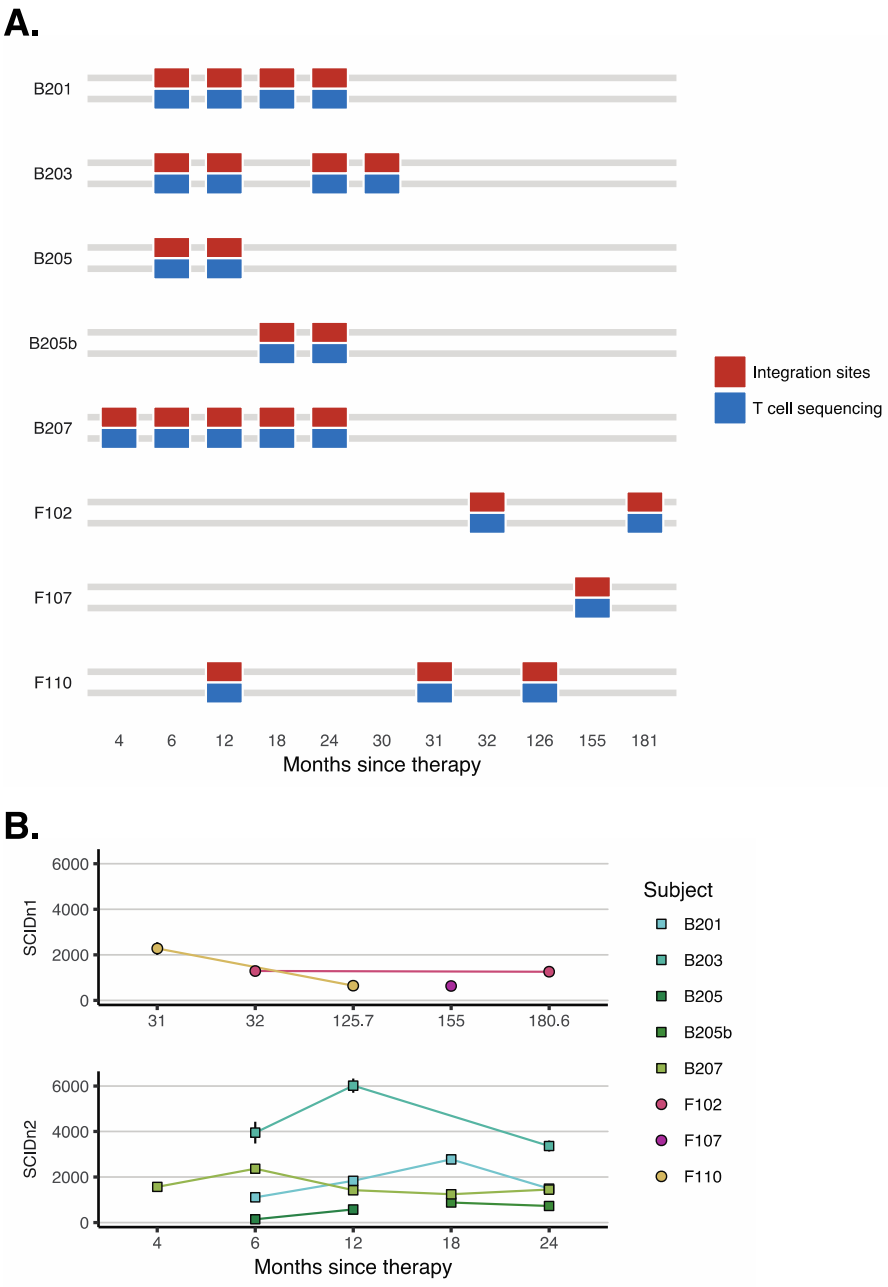


Figure 4-1.
Sampling schedule for eight SCID-X1 gene-corrected subjects studied for integration site distributions and TCRB CDR3 sequence composition. A) Times of sample acquisition after cell infusion. B) Population sizes of inferred progenitor cells deduced from marking with unique sites of vector integration. Integration site data is in Supplementary Table X, and summaries of genes near sites of vector integration are in Supplementary Figure 2. The x-axis shows the time since cell infusion. The y-axis shows the population size reconstructed using Chao1 from the numbers of unique integration sites and replicate sampling. Samples are named for the site of gene correction (B indicates Boston, F indicates France; the next digit indicates the trail 1=SCID1, 2=SCID2, and the next two digits indicate the patient number within that trial).

4.3.2. Integration site analysis

To characterize progenitor cells delivering T-cells to the periphery, we determined the sites of vector integration in patient chromosomes (31 total samples). Because of the large size of the human genome, each integration site uniquely marks the descendants of a single gene-corrected progenitor. Summaries of genes near integration sites in the most expanded clones at each time point are in Supp. Figure 4-2. The numbers of integration sites detected in purified T cell samples ranged widely, from as few as 62 to as many as 2009. Reconstruction of population sizes using the Chao1 estimator suggested minimal sizes of 144 to 6018 active progenitors.

For the SCID2 subjects, we found that the estimated total number of vector integration sites was relatively constant over the time intervals analyzed (Figure 4-1B), in the range of ~1000 predominant clones yielding circulating cells (Supp. Table 4-2). Numbers varied by subject, with subject B203 showing consistently higher levels, while B205, a case of unsuccessful reconstitution, showing consistently lower levels.

For the SCID1 subjects, numbers of unique sites identified varied from 263 to 682; reconstructed minimal population sizes ranged from 628 to 1287. Comparison of mean values shows no difference in the numbers of unique integration sites in SCID2 subjects compared to SCID1 subjects ($p=0.17$).

Population sizes of progenitors were compared between SCID1 subjects who suffered adverse events (F107, F110) versus those who did not (F102, F106), or versus SCID2

subjects. Treating time points as independent tests, no systematic differences were detected comparing within SCID1 subjects; a marginal difference was detected with a comparison of adverse events versus no adverse events (adverse event samples compared to pooled SCID1 and 2; p=0.041).

4.3.3. TCR-beta CDR3 analysis

Clonal structure of T-cell populations was investigated by analyzing TCR-beta rearrangements in genomic DNA from blood CD3+ cells (Figure 2). CDR3 region sequences were conceptually translated, and numbers of productive rearrangements quantified (Figure 4-2A). For healthy adults, numbers ranged from 18,000 to 27,000 per sample. For healthy children, numbers ranged from 18,000 to 22,000 per sample. For SCID2, most samples were close to this range (12,000 to 33,000 per sample; no significant difference in medians). For SCID1, results were slightly lower, with four out of five samples between 15,000 and 23,000. The exceptional sample was from subject F107, who suffered an adverse event and was treated by chemotherapy—in this case the number was only 5,000 unique CDR3 sequences.

For all the CDR3 samples sequenced, the numbers of CDR3 clonotypes in the subject were much larger than the numbers sequenced, so we used Chao 2 and replicate sampling to estimate the number of CDR3 variants present. Estimators are sensitive to sampling effort, so our estimates represent minimal values. Focusing on samples analyzed over multiple replicates, we estimate population sizes of 26,000 to 2,600,000 CDR3 variants. The median for SCID2 samples was not different from that of healthy children, but the

median for SCID1 was significantly lower than for the healthy children ($p=0.026$; Wilcoxon rank sum test).

Analysis of trends over time suggested that each subjects' repertoire was becoming more diverse (Figure 4-2B). The one exception was the unsuccessful case of reconstitution (B205), which did not show longitudinal increase. Inferred CDR3 population sizes in gene-corrected subjects approached or equaled sizes in healthy children, and exceeded those in healthy adults.

To begin to characterize repertoire composition, we used Bray-Curtis dissimilarity and t-SNE to cluster samples (Figure 4-2C). Replicates from each subject closely resembled those from the same subject at different time points, but differed from other subjects. No systematic differences were observed comparing SCID 1 and SCID2, or comparing either data set to healthy controls.

Recombination involving the most frequently used V (Figure 4-2D) and J (Figure 4-2E) gene segments was next quantified and compared. Usage of gene segments was quantified for healthy children and averaged, then the profile was compared to gene-corrected SCID subjects. The great majority of gene-corrected samples did not show significant differences from the distribution in healthy children, both in the analysis of V and J gene segment usage. The only exception was V gene usage in SCID2 subject B201, where particularly early time point samples were available (6 and 12 months after infusion of corrected cells). This subject was successfully corrected, so the unusual distribution is not indicative of clinical failure. We speculate that at the very early time

after gene correction newly produced T-cells are only beginning to be subjected to the homeostatic mechanisms that yield the consistent mature repertoire. Figure 2F shows the usage of V-J pairs within subjects, emphasizing the occasional outgrowth of expanded clones, potentially in response to antigen, followed by down-modulation of abundance.

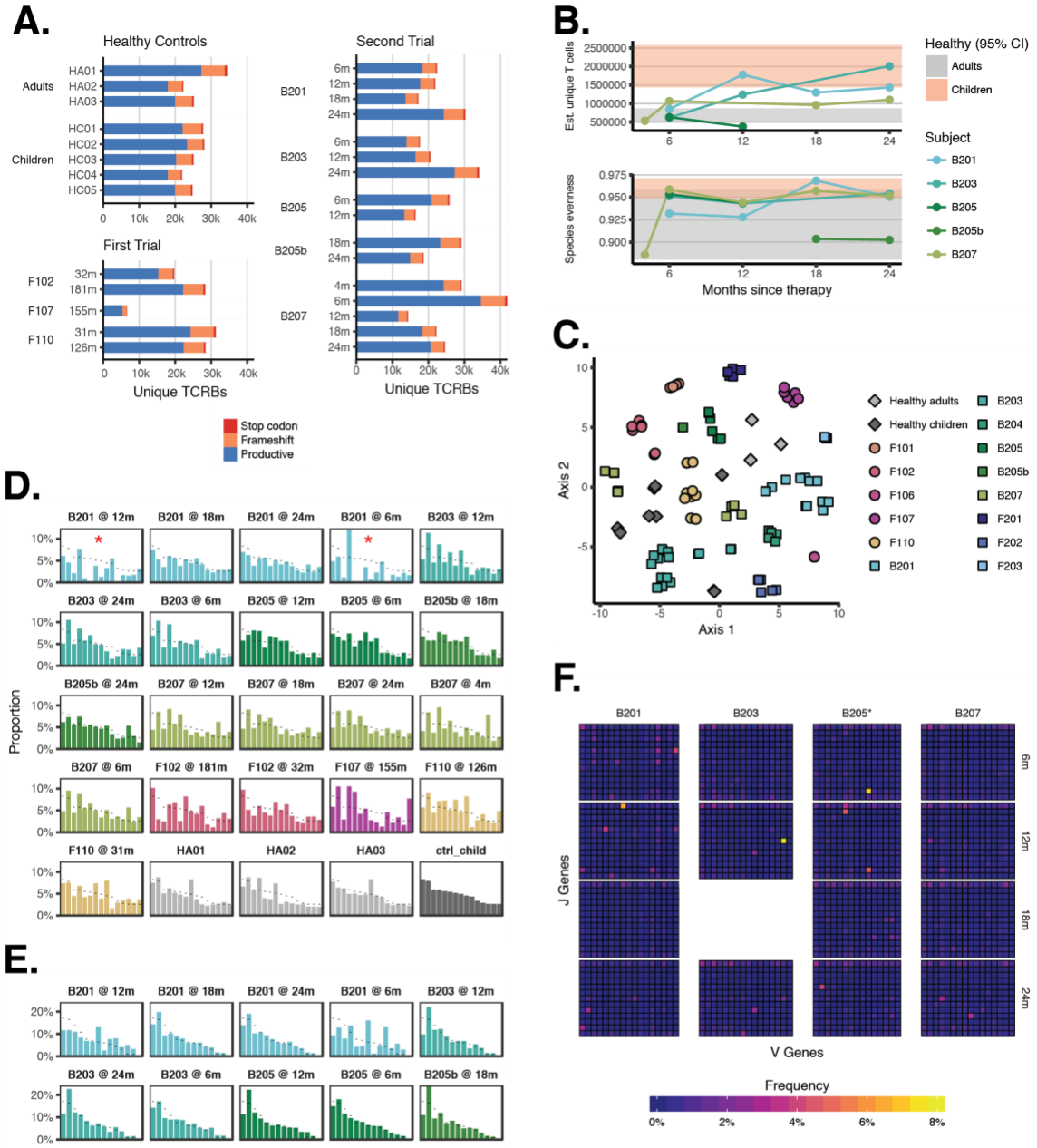

A.
Healthy Controls
Second Trial
First Trial
Adults
Children
HA01
HA02
HA03
HC01
HC02
HC03
HC04
HC05
B201
B203
B205
B205b
B207
F102
F107
F110
Unique TCRBs
Stop codon
Frameshift
Productive
B.
Est. unique T cells
Species evenness
Months since therapy
Healthy (95% CI)
Adults
Children
Subject
B201
B203
B205
B205b
B207
C.
Axis 1
Axis 2
Healthy adults
Healthy children
F101
F102
F106
F107
F110
B201
B203
B204
B205
B205b
B207
F201
F202
F203
D.
Proportion
E.
Proportion
F.
J Genes
V Genes
Frequency

Figure 4-2.
Analysis of TCRB CDR3 sequences. A) The unique numbers of rearranged genes detected are shown. The colors indicate in frame rearrangements (blue), frameshifts (tan) and stop codons (red). B) Richness and evenness of the inferred TCRB CDR3 populations. Patients are color coded as indicated on the right. The ranges of healthy adults and children are shown by the grey and salmon shading, respectively. C) Clustering of the samples sequenced using Bray-Curtis similarity and t-SNE. The association of patients with samples is shown by the key at the right. D) V gene usage. The patient of origin is marked at the top of each panel. E) J gene usage. F) Heat map summarizing the frequencies of utilization of the most common V and J pairs. Subjects studied are marked at the top. Time of sampling is shown on the right.

4.3.4. Tracking T-cell ontogeny

Comparison of the estimated lower bound for the number of unique integration sites per sample with the lower bound for the number of unique TCR sequences allows estimation of a lower bound on the number of cell divisions required to generate the TCR-beta cell population from gene-corrected precursors (Figure 4-3. We calculated the minimum number of cell divisions as the base-2 logarithm of the difference between the estimated population size of unique T cells and the estimated population size of integration sites:

$$\text{CellDivisions} = \log_2(\text{TCRs} - \text{IntSites})$$

We found a relatively consistent range of minimum cell division values across patients and timepoints, with a median of 8.41 and a minimum and maximum of 5.97 and 17.45. Considering only the patients with successful, non-repeated therapy (B201, B203, and B207), the range was tighter, with a median of 8.41 and a minimum and maximum of 5.97 and 14.72, respectively. The fraction of progenitor cells that die in the thymus is unknown, so our estimates are lower bounds for the required number of cell divisions.

The highest values were for subject B205, for whom reconstitution was unsuccessful. We speculate that homeostatic mechanisms may have resulted in signaling to a limited number of progenitors to divide with increased frequency in this subject.

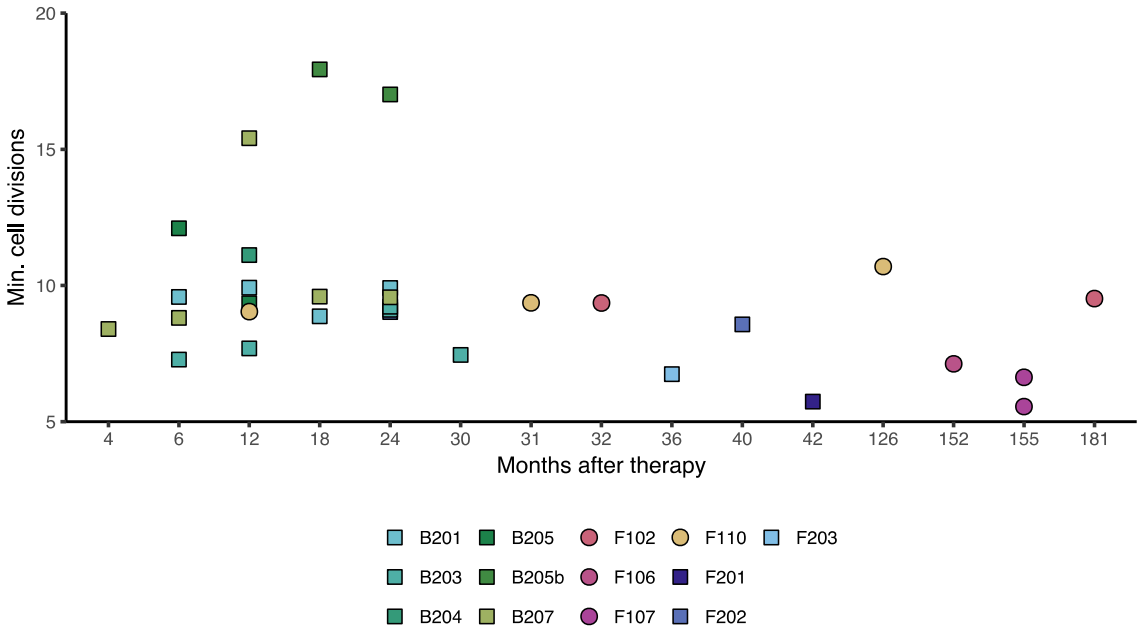


Figure 4-3.
Minimum numbers of cell divisions between progenitors and daughter T cells. The x-axis shows time after corrected cell infusion. The y-axis shows the estimated number of cell divisions calculated as described in the text. The subjects studied are indicated beneath the figure as indicated by the color code.

4.3.5. Response of the microbiome to reconstitution

Microbiota community structure was analyzed for six gene-corrected subjects (Figure 4-4A) by extraction of DNA from swabs (oral and nasal samples) or stool (fecal samples) followed by shotgun metagenomic sequencing. Numbers of samples available per subject ranged from one to seven. The numbers of sequences acquired per sample averaged 3,625,000 (oral), 72,000 (nasal) and 8,746,000 (fecal). Sequencing reads were quality filtered as described in Methods and then assigned to microbial taxa using Kraken (Wood & Salzberg, 2014) .

The analysis of gut microbiota (Figure 4-4B) emphasized the differences between gut microbiota of healthy children (left four columns) versus SCID gene-corrected subjects. The healthy subjects were colonized predominantly with Bacteriodes, which is typical of healthy gut, whereas the SCID-subjects showed a range of major colonists. B201 was colonized mainly with *Bifidobacteria*, which is characteristic of healthy breast-fed babies (Pannaraj et al., 2017). However, subjects B203 and B205 (early times) were colonized with Veillonella, which is typically an oral bacteria, possibly indicative of colonization with this organism along the length of the GI tract. Subject B204 showed high level colonization with Enterobacteriaceae, typical of dysbiotic states. Early samples from subjects B207 and F201 were dominated by viruses, even though whole stool was sequenced, with infection by adenovirus and bocavirus respectively.

Oral samples from the healthy controls were dominated by typical oral bacteria, including *Prevotella*, *Streptococcus*, *Neisseria*, and *Haemophilus*. In contrast, B203 at the earliest time point was dominated by Bocavirus, B204 and B205 showed high level domination by Streptococcus, and B207 was dominated by *Rothia*.

In nasopharyngeal samples, healthy subjects were dominated by *Moraxella*, *Staphylococcus* and *Propionibacterium*. SCID subjects contained these lineages to varying degrees, but also showed high level colonization by *Streptococcus*, *Corynebacterium*, and others.

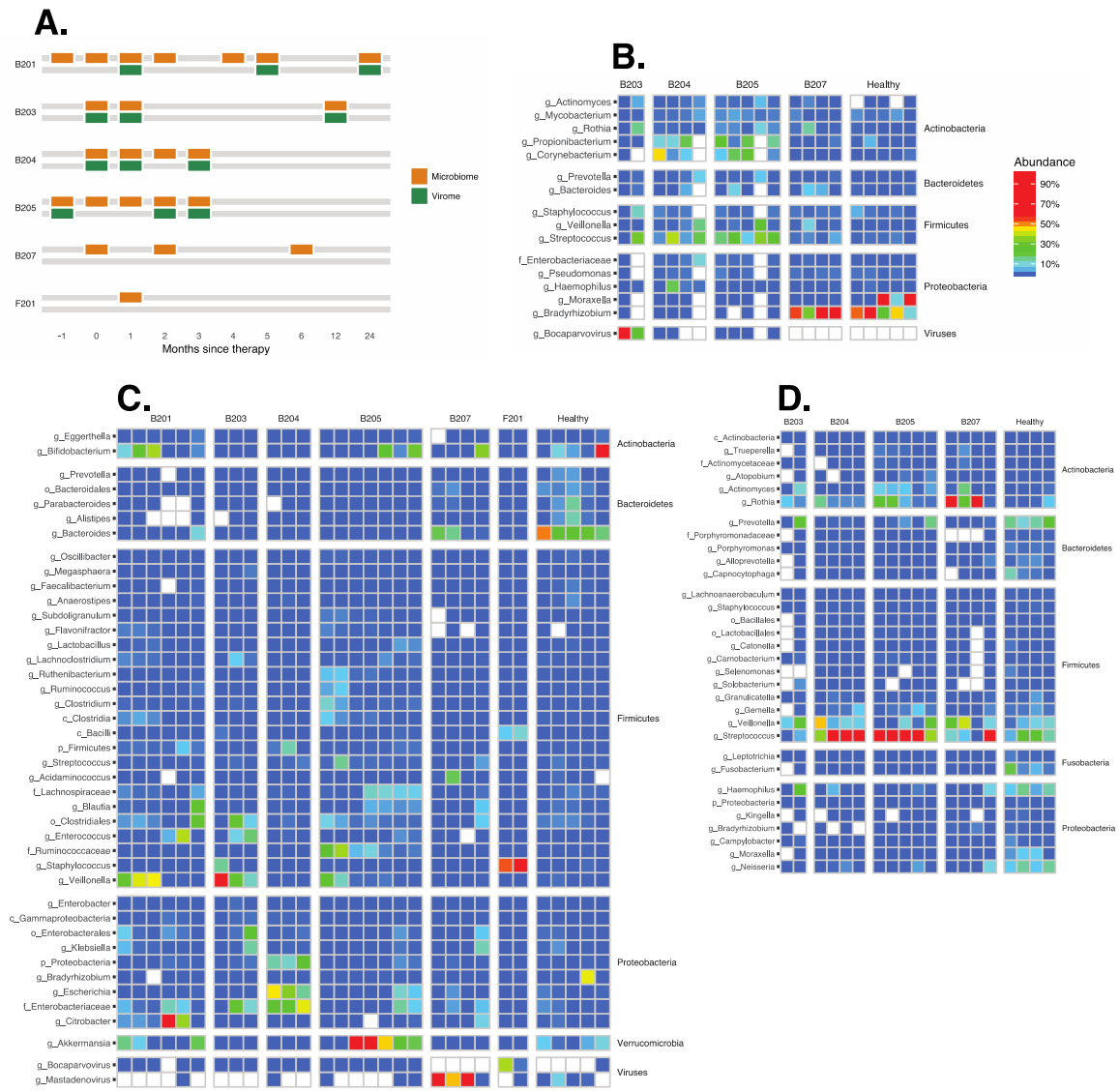


Figure 4-4.
Longitudinal analysis of the microbiome during SCID-X1 gene correction. A) Timing of sample acquisition. B) Longitudinal analysis of the nasopharyngeal microbiome. Each column indicates a sample. Samples are grouped by subject as indicated at the top. Each row summarizes the proportions of a specific microbial taxa inferred using Kraken. Abundance is color coded as indicated to the right. C) As in B, but stool samples. D) As in B, but oropharyngeal samples.

We compared the full microbial compositions of the samples using Bray-Curtis dissimilarities and found that the three sample types clustered by body site of origin (Figure 4-5A, PERMANOVA p-value < 0.001). We also found that each subject

clustered with themselves throughout all timepoints, reflecting consistently higher inter-subject variability than intra-subject variability (PERMANOVA p-value < 0.001). In several patients (B201, B205, and B207), the microbial communities began resembling healthy children more at late times after cell infusion (Figure 4-5B), potentially reflecting a combination of improving immune function and reduction in antibiotic usage.

Analysis of taxonomic richness provides insight into microbial community health, since low richness is often associated with abnormal outgrowth of opportunistic organisms. For stool, richness increased for 4/6 subjects over time, potentially indicative of improving gut health (Figure 4-5C). Unexpectedly, of the four subjects for whom oral samples were available, all showed abnormally low richness at every time point. For many this was associated with particularly high Streptococcus colonization, an observation that might be of interest to investigate further for possible clinical implications. Nasopharyngeal microbiota showed richness comparable to healthy controls.

SCID patients were treated with a wide variety of antibiotics before and after therapy to mitigate opportunistic infections. To assess whether this was associated with an increase in antibiotic resistance gene representation, we quantified antibiotic resistance genes in the stool sequence samples using ShortBRED and the CARD antibiotic resistant factor database (Figure 4-5D). We saw an increase in the total quantity of antibiotic resistance genes. The SCID patients had a higher level of antibiotic resistance genes in their stool than healthy subjects (pooled across timepoints, $p = 0.048$). These results indicate that the

antibiotic regimen increases both general levels of antibiotic resistance genes. For patient B201, who had a timepoint 24 months after therapy, the level of antibiotic resistance genes decreased substantially between the 5 month and 24 month timepoint. At this timepoint, the patient was no longer on any medications, which suggests that antibiotic resistance gene load may decrease after treatment ceases.

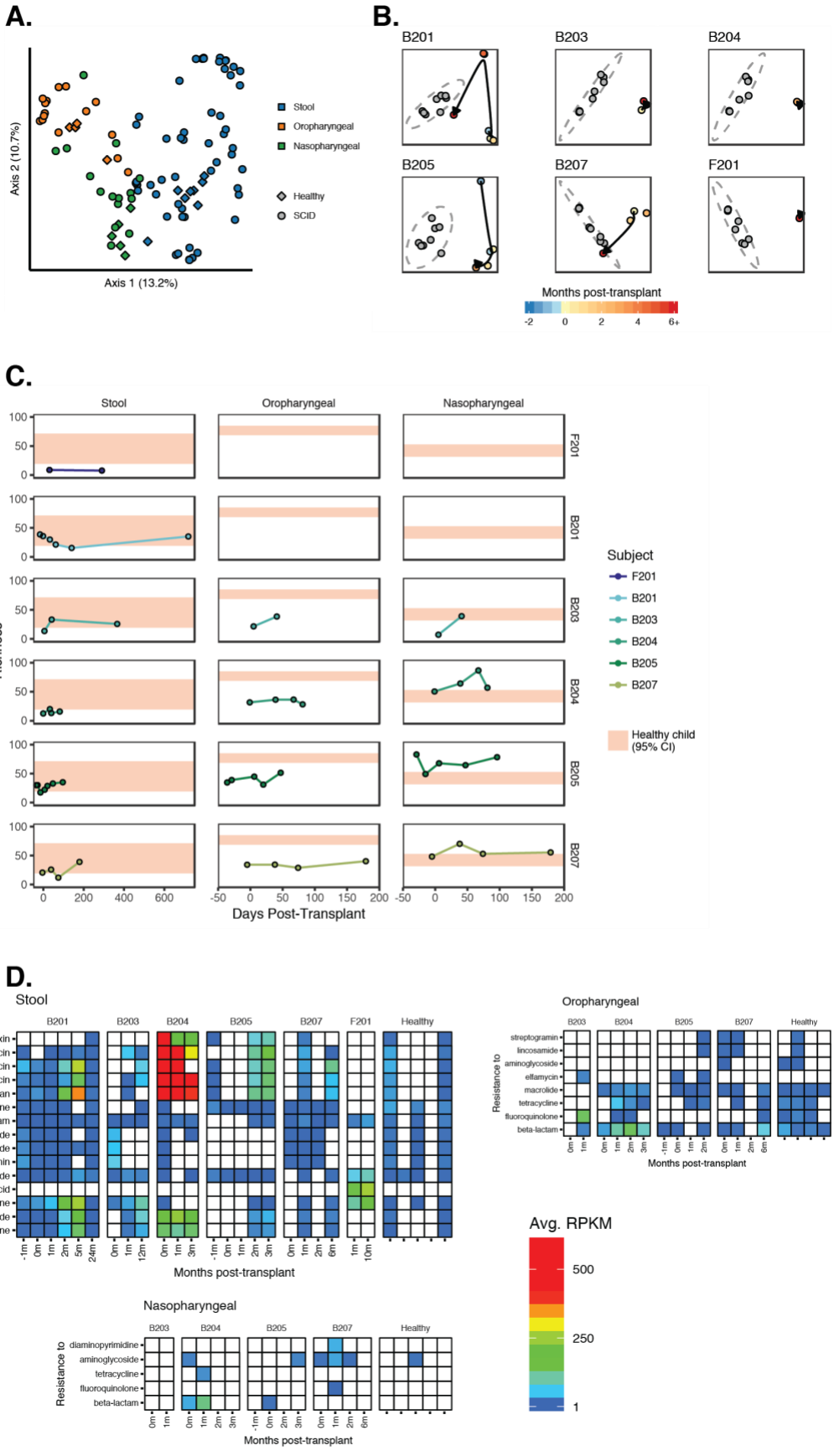

A.
Axis 2 (10.7%)
Axis 1 (13.2%)
Stool
Oropharyngeal
Nasopharyngeal
Healthy
SCID
B.
B201
B203
B204
B205
B207
F201
Months post-transplant
-2
0
2
4
6+
C.
Stool
Oropharyngeal
Nasopharyngeal
Richness
Days Post-Transplant
Subject
F201
B201
B203
B204
B205
B207
Healthy child (95% CI)
D.
Stool
Resistance to
polymyxin
elfamycin
rifamycin
fosfomycin
triclosan
tetracycline
beta-lactam
macrolide
lincosamide
streptogramin
aminoglycoside
fusidic acid
diaminopyrimidine
sulfonamide
fluoroquinolone
Months post-transplant
Oropharyngeal
streptogramin
lincosamide
aminoglycoside
elfamycin
macrolide
tetracycline
fluoroquinolone
beta-lactam
Nasopharyngeal
Avg. RPKM
500
250
1

Figure 4-5.
Microbiome community analysis. A) Sample clustering using Bray-Curtis dissimilarity. Different sample types are marked by the colors, healthy versus SCID are shown by the shapes. B) Comparison of stool samples for each patient queried to healthy controls. Samples were clustered using Bray-Curtis dissimilarity. Each panel compares one SCID subject samples (indicated at the top) to healthy control samples (shown in grey). Elapsed time is shown using the color code (bottom). C) Representation of selected antibiotic resistance genes in the three sample types studied. Each column indicates a metagenomic data set from the subject listed at the top. Each row summarizes the abundance of an antibiotic resistance gene class. The tiles are colored by reads per kilobase of target per million sequences reads (RPKM); the color code is to the right of the panel.

4.3.6. Response of the virome

Viral particles were partially purified from stool, the DNA and RNA purified and sequenced. Reads were assigned using Kraken, then extensive filtering was carried out to remove contaminants and artifacts. Figure 4-6 shows the resulting attributions. For RNA viruses, high level colonization was detected with astrovirus (B204 and B205) and sapovirus (B203). For DNA viruses, numerous bacteriophage lineages were seen. For viruses infecting animal cells, high levels of infection were seen for adenovirus in B201, bocavirus in B203, and betatorquetenovirus (anellovirus) in B205. The adenovirus and bocavirus infections diminished over time after successful gene correction. Betatorquetenovirus is a normal commensal, and persisted.

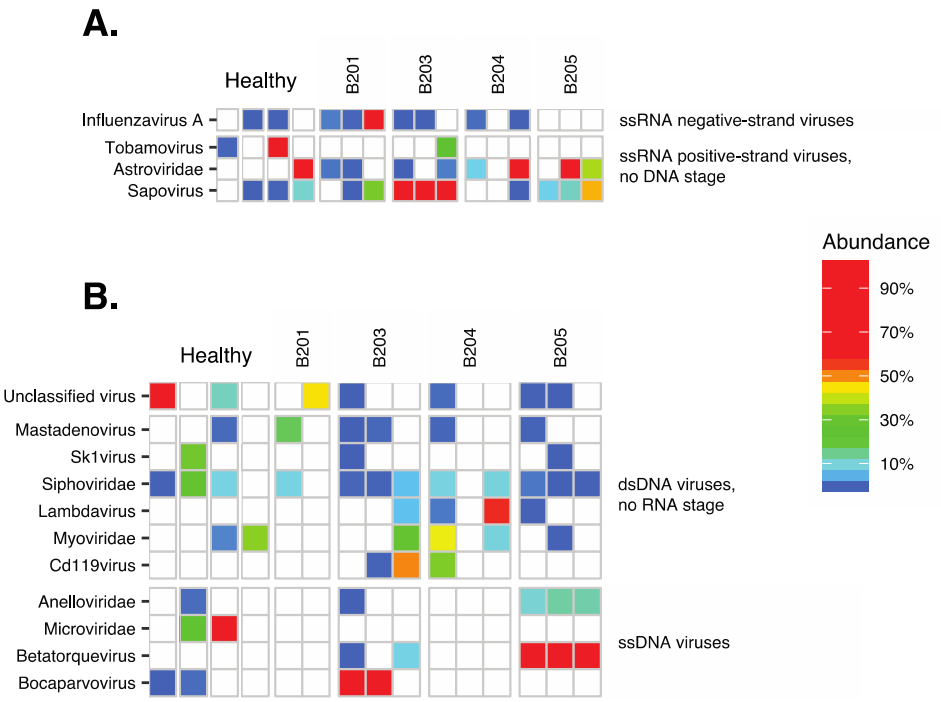


Figure 4-6.
Virome analysis. A) Heat map summarizing RNA viruses detected. Each column indicates a sample from the patient indicated at the top of the heatmap, each row indicates a type of virus. The tiles are colored according to proportion of total viral reads. B) Heat map summarizing DNA viruses detected. Markings as in A.

4.4. Discussion

We present here a first look at the co-development of the microbiome and immune system in patients after gene therapy for SCID. We used targeted sequencing of vector integration sites to model the number of gene-corrected progenitor cells, and TCR sequencing to capture the development of the T cell repertoire following therapy. The combination of these data allowed a lower-bound approximation of the number of cell divisions required to progress from a lymphopoetic stem cell to a circulating T cell, a measurement uniquely possible thanks to gene marking by integration. We used shotgun metagenomic sequencing in longitudinal samples to track early-term changes in the gut,

oral and nasopharyngeal microbiome coincident with therapy. In a subset of these timepoints, we purified and sequenced RNA and DNA viruses to show clearance of pathogenic viruses with immune reconstitution and, in some patients, the presence of viruses not reported in clinical records.

This estimate of cell divisions is limited by wide confidence intervals in the estimators for unique TCRs and integration site population sizes, but despite that most of the timepoints indicated roughly nine cell divisions between the progenitor stem cell and a circulating T cell. This is a minimum value because the amount of cell death during T-cell ontogeny is unknown and likely high. This helps specify the functional capacity of the lymphoid precursor cell targeted in initial transduction.

The microbiome analysis provided new evidence on how SCID gene therapy patients progress to healthier phenotypes after successful therapy. Their immune cell diversity, as measured by TCR sequencing, consistently moved from the oligoclonal range of older adults and into a higher diversity range more characteristic of healthy children of that age. Further, their overall circulating T cell richness showed a similar movement towards a healthy and age-typical state. This is concordant with the observed results of healthy immune function in these patients and shows that sequencing metrics of the T cell repertoire correlate with clinical outcomes.

As in their immune repertoires, the microbiota of these patients changed in ways that resembled the microbiomes of healthy children. We saw an increase in diversity and richness of the microbes in the gut and oro- and nasopharygeal compartments and the

outgrowth of bacteria associated with healthy gut function. In some patients, we were also able to detect bacteria that may be associated with negative outcomes and antibiotic resistance, potentially as a result of extended exposure to antibiotics and hospitalization.

An unexpected finding was the consistently low diversity of SCID oral samples, and the high colonization with Streptococcus. It will be of interest to assess whether this is seen in other immunodeficient subjects, and target investigations of possible oral pathology associated with this organism.

In some of these patients, we found viruses known to cause enteropathic conditions in humans, including astrovirus, bocavirus and adenovirus. In the case of patient B207, the adenovirus detection was clinically corroborated with a diagnosis of adenoviremia. In both, the viruses were cleared in later timepoints, likely as a result in part the newly-functioning immune system.

This study is also limited in a number of important ways. Most importantly, our sample size was quite small. Moreover, clinical considerations often led to inconsistent microbiome sampling based on patient and clinician availability, as sample collection was a secondary concern to patient monitoring. As a result, our ability to make broad inferences based on this sample set is limited.

Opportunities to study similar systems of immune system/microbiome dynamics are present in SCID patients who undergo hematopoetic stem cell transplantation (HSCT). A pilot study looking at the microbiome development in SCID patients after HSCT (Lane,

2015) was also performed, though with 16S rRNA gene amplicon sequencing rather than shotgun sequencing, no healthy controls, shorter sampling period, and no concurrent TCR sequencing. They found inconsistent results in microbiome diversity after transplantation, possibly due to the limited temporal range of the study, but did note that there were clear differences in the microbiota before and after transplantation in the four subjects they studied.

In summary, these data illustrate some of the uses of multi-omic data in assessing outcome in human gene therapy. As more of these studies are carried out, it will be possible to more fully assess the utility of such data. Of particular interest will be any signatures that help forecast outcome and provide new opportunities for initiating specific interventions.

4.5. Methods

4.5.1. Human subjects

Patients were recruited as described (M. Cavazzana-Calvo et al., 2000; Hacein-Bey-Abina et al., 2014). We collected the same sample types from six healthy children between the ages of 21-43 months under IRB 13-010072. We obtained sorted CD3+ T cells from three anonymous healthy adult donors above the age of 18 from the Human Immunology Core at the University of Pennsylvania. All samples were stored at -80^{o}C.

4.5.2. Integration site analytical methods

Integration site sequences were determined using two different methods due to changes in technology over the period of patient monitoring. In the first, 454/Roche pyrosequencing was used to determine integration site placement (G. P. Wang et al., 2007). In the second, Illumina paired end sequencing was used (C. C. Berry et al., 2012; C. C. Berry et al., 2017; Hacein-Bey Abina et al., 2015; Sherman et al., 2017). In both, DNA was broken using shearing or restriction enzyme cleavage, then DNA adaptors ligated on to the broken DNA ends. Nested PCR was then used to amplify from the linker to the integrated vector, and the intervening segment of human DNA sequenced. All integration site sequence analysis was be carried out in quadruplicate to minimize PCR jackpotting. All sample sets were worked up together with human DNA lacking integrated lentiviral sequences to monitor for PCR contamination, which was typically undetectable. Different linkers were be used for ligation-mediated PCR for each sample in a set to block PCR cross over. All samples were bar coded on both ends of the molecule, and only those with correct bar code pairs analyzed, thereby suppressing artifactual molecules resulting from PCR recombination. A total of 31 samples were analyzed, yielding a total of 24,170 integration sites.

4.5.3. TCR sequence analysis

TCR sequencing was performed on whole blood samples that had been fractionated to yied T cell (CD3+) or PBMC fractions. Genomic DNA was isolated and sequenced at Adaptive Biotechnologies to determine CDR3 region sequences of the TCR beta locus

(H. Robins et al., 2012). Data were analyzed using the immunoSeq Analyzer version 3.0. A total of 40 samples were analyzed, yielding a total of 32 million TCRB sequences.

4.5.4. Microbiome sequencing

DNA was isolated from fecal, oral and nares samples using the following procedures: Small aliquots of fecal material (<= 1ml) and the tips of each swab (for oral and nares swabs) were deposited into a PowerSoil bead tube. DNA was extracted using standard the MoBio PowerSoil DNA extraction protocol, with one or more blank extraction controls worked up simultaneously with each set of samples. Work spaces were decontaminated using bleach and UV irradiation. The resulting DNA from all samples and blank controls was sequenced on an Illumina HiSeq 2500 using NextSeq chemistry and standard Illumina dual barcoding for each sample.

Due to artifacts in genomic sequences for members of the Apicomplexa family, many reads from a common water contaminant (*Bradyrhizobium*) were cross-annotated as belonging to Apicomplexa. Consequently we removed all *Bradyrhizobium* and Apicomplexa reads before further analysis due to their uncertain provenance.

4.5.5. Virome analytical methods

Viral particles were isolated from a subset of the fecal samples using a protocol adapted from (Minot et al., 2013). Fecal samples were homogenized and filtered through a 0.4 micron filter. The filtered samples were then treated with DNaseI and RNAseI to remove

exogenous nucleic acids. Combined nucleic acids were then purified from the sample using the Qiagen UltraSens Virus Kit.

To obtain DNA viruses, we amplified viral genomes in an aliquot of the combined nucleic acids using the Illustra Genomiphi V2 DNA amplification kit. Resulting amplified DNA was quantified using PicoGreen and stored at -20°C. To obtain RNA viruses, we treated a separate aliquot of the combined nucleic acids with DNAse+ and then performed reverse transcription of the RNA to cDNA using the SuperScriptIII First-Strand Synthesis System from Life Technologies and second strand synthesis using Sequenase. The resulting cDNA was quantified with PicoGreen and stored at -20°C.

Resulting DNA was prepared and sequenced using NextSeq for library preparation and an Illumina HiSeq 2500 for sequence acquisition. The same postprocessing pipeline was used, including all quality-control, host removal, and read annotation steps. For analysis, we considered only reads that fell under the Virus classification, and removed the following viral annotations as being reagent contamination (Abbas et al., 2016; Clarke et al., 2017): Enterobacteria phage M13, Enterobacteria phage T7, Enterobacteria phage phiX-174 sensu lato, Bacillus phage phi29, and Pseudomonas phage phi6, human herpesvirus 6 and 7, and Shamonda virus.

4.5.6. Bioinformatic methods

To estimate population sizes of T cell progenitors from integration sites we used a jackknifed Chao1 estimator (abundance-based), and for TCR sequence data, we used the incidence-based Chao2 estimator (Chao, 1987).

Sequence reads for all metagenomic samples were processed using the Sunbeam pipeline (https://github.com/eclarke/sunbeam). Reads were quality-controlled by trimming low-quality bases and adapter sequences using Trimmomatic (Bolger et al., 2014), and host reads were removed using BWA (Li & Durbin, 2010). The remaining reads were filtered for low-complexity sequences using dustmasker (Camacho et al., 2009) and Komplexity (https://github.com/eclarke/komplexity). The reads that remained were assigned taxonomy via Kraken (Wood & Salzberg, 2014) and a custom database built on all microbial genomic sequences in RefSeq release 79 (O'Leary et al., 2016).

Antibiotic resistance gene levels were assessed using ShortBRED (Kaminski et al., 2015) and a marker gene database built from CARD (Jia et al., 2017) available on the ShortBRED website (https://bitbucket.org/biobakery/shortbred/downloads/ShortBRED_CARD_2017_markers.faa.gz). The same reads used as input to the Kraken taxonomic classifier were used as input to ShortBRED.

Final analysis and figure generation was performed using the R statistical software (R Core Team, 2017). Bray-Curtis dissimilarity and other ecological metrics were calculated

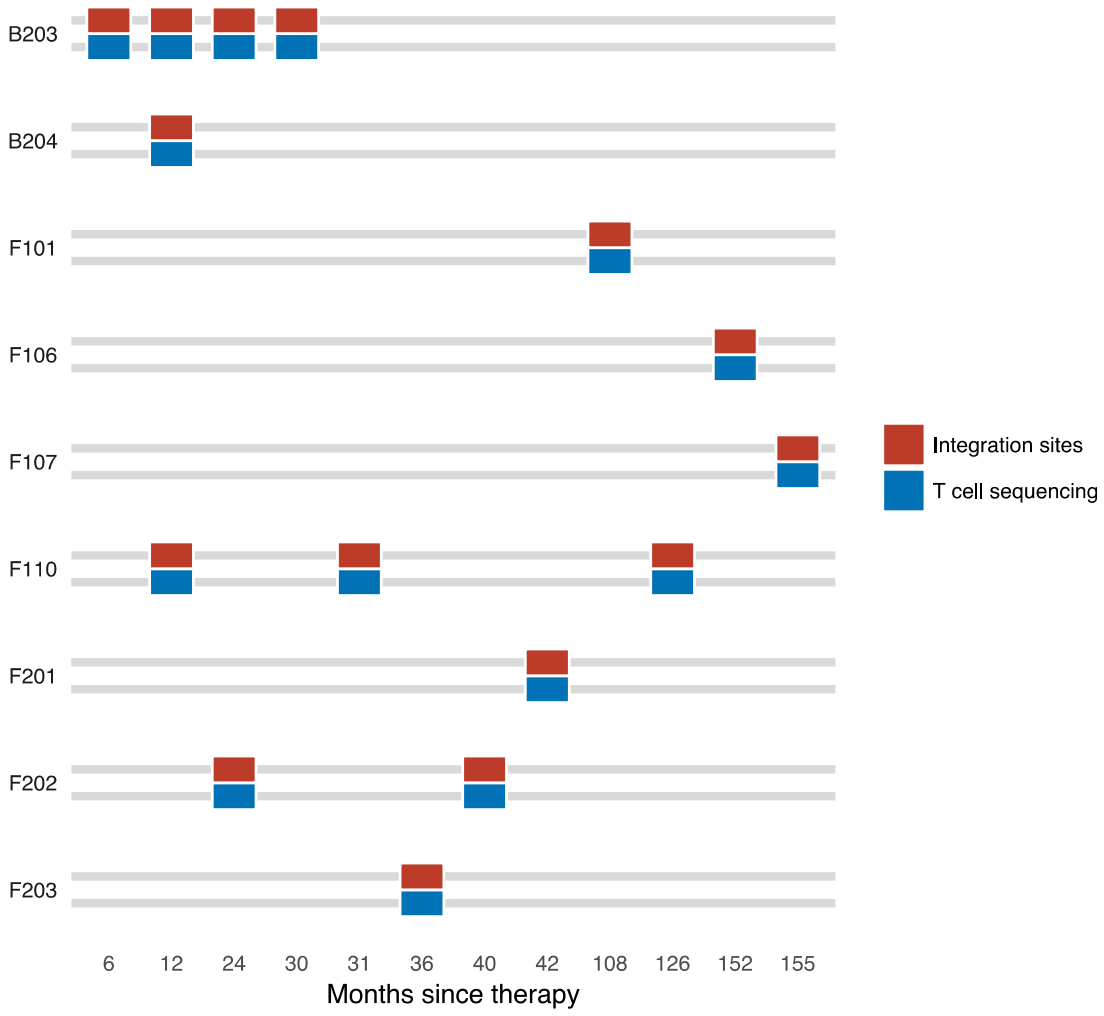
using the R package vegan (Dixon, 2003). The full code listing and post-processed data used for analysis and figures is available online at https://github.com/eclarke/scid-multiomics-paper.

4.6. Acknowledgements

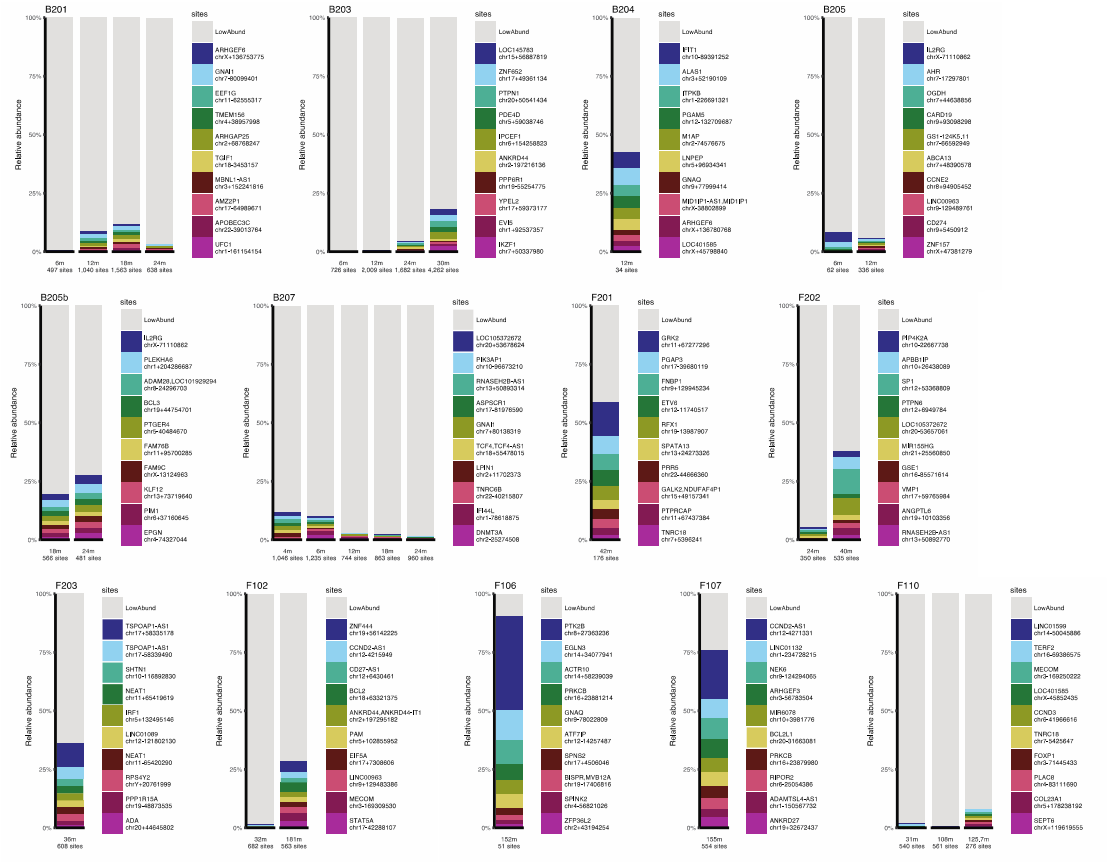
We are grateful to members of the Bushman laboratory for help and suggestions. This work was supported by AI 052845-13, AI 082020-05A1, AI 045008-15, U19AI117950-01, UMIAI126620 to F. D. B., the Penn Center for AIDS Research, the Penn Human Immunology Core (P30-CA016520) and the PennCHOP Microbiome Program.

4.7. Supplemental Material

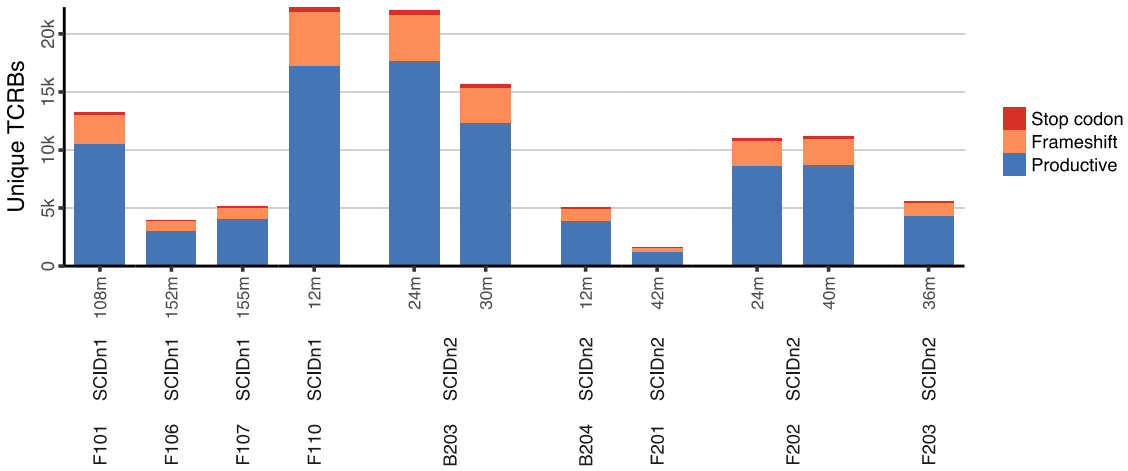
4.7.1. Supplemental Figures



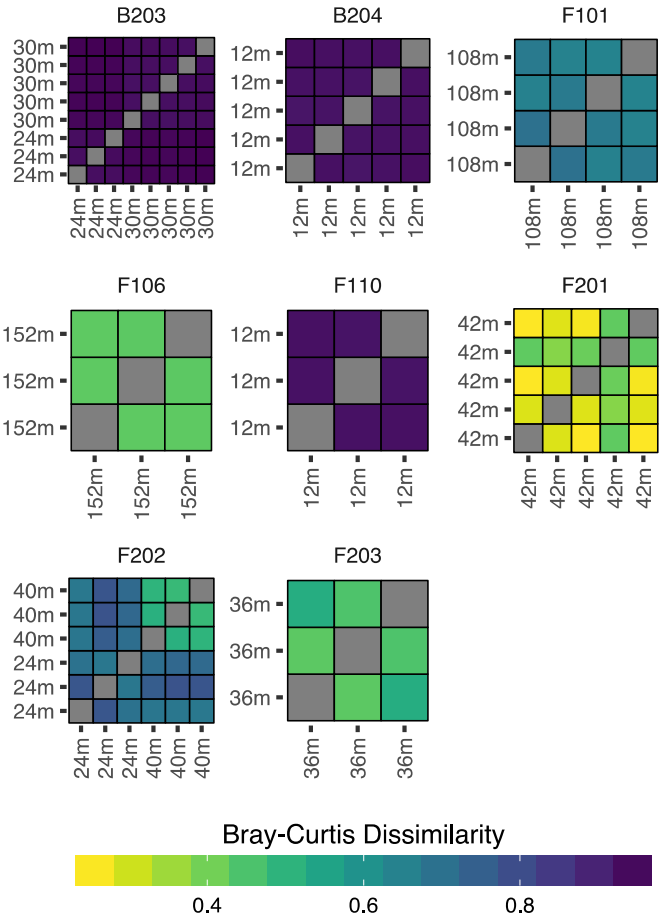
Supp. Figure 4-1. PBMC samples timeline.



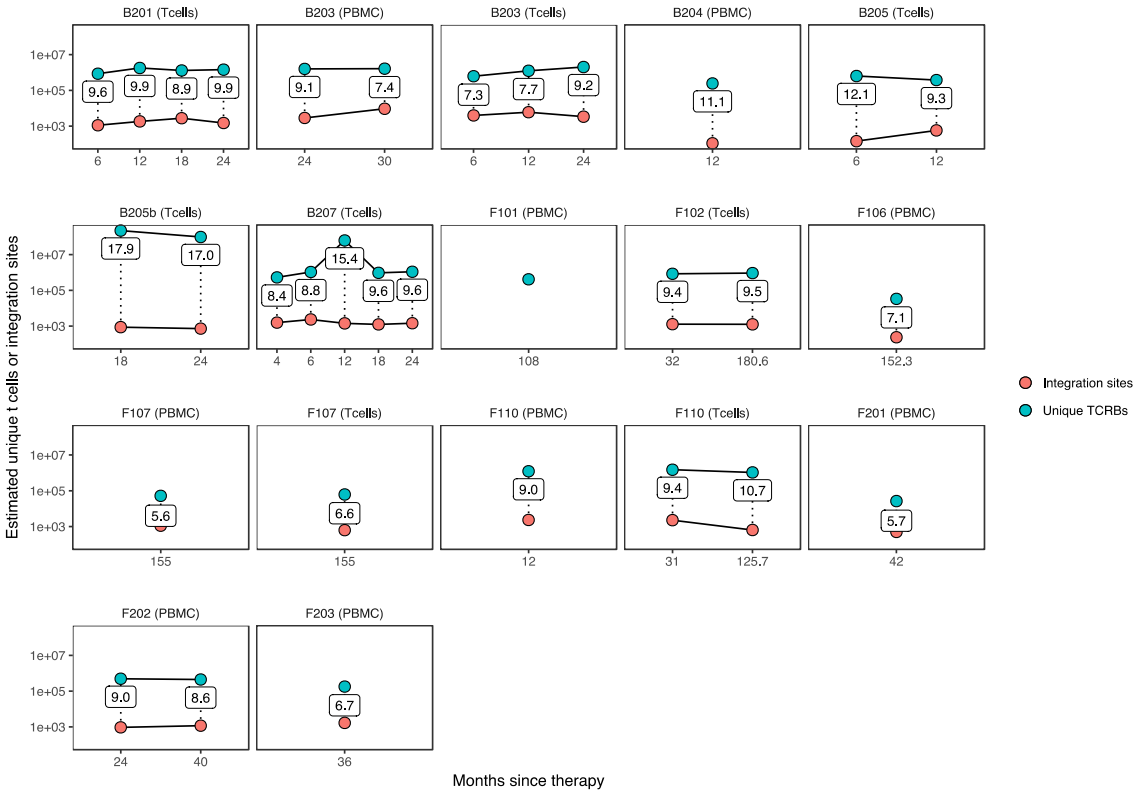
Supp. Figure 4-2. Genes near integration sites.
Stacked bar graphs summarizing genes at or near integration sites for each subject and sample studied here.



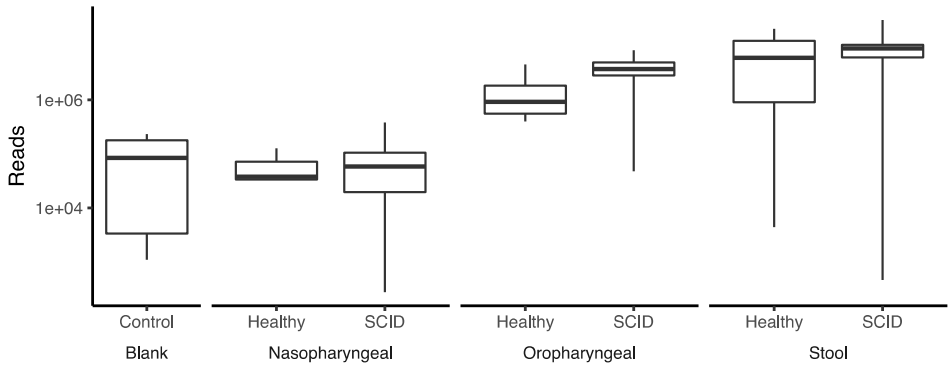
Supp. Figure 4-3. TCRB population characteristics (PBMC).
Stacked bar graphs summarizing characteristics of the TCRB repertoire sequenced from PBMC cells.



Supp. Figure 4-4. TCRB repertoire similarity (PBMC).
Bray-Curtis dissimilarity measurements for the TCRB repertoires sequenced from PBMC cells.



Supp. Figure 4-5. Lymphocyte progenitor cell divisions.
Population size estimates for progenitors and daughter t cells, with the minimum number of doublings (cell divisions) required.



Supp. Figure 4-6. Mean read counts for metagenomic sequencing.

4.7.2. Supplemental Tables

All supplemental tables are attached in the Digital Supplement named SupplementalTables.xlsx.

Supp. Table 4-1. Subjects in this study.

Supp. Table 4-2. Vector integration site data.

Supp. Table 4-3. TCR-beta repertoire data.

Supp. Table 4-4. Microbiome data

Chapter 5. Conclusions and Future Directions

In this thesis I describe novel approaches to using metagenomic sequencing in idiopathic and immune-related diseases. Metagenomics has found fruitful application in studies involving gastrointestinal disorders and dysbiosis of existing communities. However, the utility of metagenomics has been less clear in other types of diseases. The heightened sensitivity of next-generation sequencing at retrieving rare microbes will be key to uncovering causal pathogens in idiopathic diseases, but standard approaches suffer from extreme sensitivity to contamination and false-positives (Lauder 2016). Meanwhile, the interactions between microbiota and the immune system are beginning to be more clearly understood (Arpaia 2013, Kamada 2014, Atarashi 2013) but their interplay in immune disorders is less clearly defined.

In Chapter 2, I present a study where we looked for microbial triggers of sarcoidosis, a granulomatous disease with no known cause. Our effort involved experimental and statistical methods to rigorously account for environmental contamination. We collected paired environmental controls for every sample and specific for every sample type: for formalin-fixed, paraffin-embedded (FFPE) tissue we sampled the surrounding blank paraffin, while for bronchoalveolar lavage (BAL) we sampled the prewash rinse of the scope done prior to the procedure. These paired environmental controls were integrated into the differential abundance testing by the use of a generalized linear mixed model (GLMM) in which the background level of a taxa was allowed to vary in each sample pair.

This approach demonstrated exceptional precision in the face of strong environmental covariates. We found that in one set of FFPE tissue samples, high levels of *Aspergillus* fungi were present in sarcoidosis samples and not in healthy controls. But this same difference was seen in the paired environmental controls, indicating that *Aspergillus* was environmentally derived. (We later traced the source of this contamination back to difference in storage sites between the two sample groups.) In studies without this experimental design, we would have been unable to disentangle the environmental confounder from the study group and spent significant time investigating *Aspergillus* fungi as potential sarcoidosis triggers. However, differential abundance testing with the GLMM did not identify *Aspergillus* as enriched. Rather, it identified a different fungus in the Cladosporiaceae family as being enriched in sarcoidosis tissue over both healthy tissue and the background environment. Cladosporiaceae fungi, while not known to be pathogenic, are significant allergens and can cause disorders including hypersensitivity pneumonitis (Chiba et al., 2009; Silva & Ekizlerian, 1985), so this is a biologically plausible trigger in accordance with Relman's postulates (Fredricks & Relman, 1996).

This paired environmental control study design and associated model are applicable to other low-biomass metagenomic studies that aim to find differences between two groups. While consensus has been reached on the necessity of blank controls in these studies to characterize the input from reagents (Kim et al., 2017; Salter et al., 2014), those controls alone are not always sufficient–for instance, they would not have uncovered the *Aspergillus* contaminant as described above. There exist other approaches for determining the source of microbial signatures in a sample (Knights et al., 2011), but this

is the first able to translate these contaminating sources into terms that affect differential abundance testing. Many clinical studies would benefit from this approach, especially those that investigate microbial communities in sources previously considered sterile in healthy individuals, such as the brain or placenta. In these studies, collection and storage methods may contribute significant amounts of microbial biomass as a percentage of the total. Without the controls, researchers may be erroneously reporting the presence of endogenous or enriched microbiota when in fact they are subtle contaminants.

In Chapter 3 I describe a computational method that efficiently designs primers to selectively amplify a target's genome from a complex mixture. This program, *swga*, makes the method described in (Leichy 2014) significantly more accessible to a broad range of researchers. The wet-side method, selective whole-genome amplification (SWGA), uses a set of phi29 primers that bind more frequently to a target's genome than the background to preferentially amplify the target using multiple-displacement amplification. With the correct primers, the resulting amplicon will have much higher amounts of target DNA proportionally than the starting mixture. Consequently, less sequencing effort is required to achieve high depth-of-coverage of the target genome. The *swga* program described in this chapter makes it possible to rapidly design primer sets for this assay, which was originally a manual and error-prone process. The program works by first identifying all *k*-mers of a range of lengths in the target and background genome. It then filters the *k*-mers for undesirable characteristics including self-complementarity and suboptimal melting temperature range, or uneven binding on the target genome. It then assembles a graph representation with primers as vertices, and compatible primers–

i.e. those that do not form primer dimers–are linked by edges. Each primer is weighted by how frequently it binds the background genome. The program then searches for cliques, or completely-interconnected subgraphs, with the smallest possible weight. These cliques form the basis for selective primer sets for the SWGA method.

In this chapter, I demonstrated the efficacy of this program and approach on three real-life host/parasite systems. SWGA, when used with a well-chosen primer set, significantly reduces the costs associated with repeatedly obtaining high-coverage genomes of a given target. This is especially critical when retrieving genomes from targets that are difficult or impossible to culture, especially in the context of pathogen or parasite genomics. In this sense it is complementary to our efforts in Chapter 2 to identify microbial agents in sarcoidosis: one possible use for SWGA is the isolation and characterization of the precise Cladosporiaceae fungi in the sarcoidosis samples to see if there are strain-level differences that may be causing disease symptoms.

In Chapter 4, I presented our efforts to characterize the interactions between the microbiome and immune system of SCID patients after gene therapy. The gene therapy trials for SCID provided a unique opportunity to observe a cohort of humans born without an immune system, but with colonized microbiomes (normally, the microbiome and immune systems co-develop in healthy babies). Once the gene therapy was administered, these patients' immune systems began to develop, and we were able to track how the immune system and microbiome in these patients changed in response to each other. To do this, we used T cell repertoire sequencing and metagenomic shotgun

sequencing on a longitudinal sample set. We found that patients generally reconstituted a normal-looking immune system when the therapy worked, which was concordant with reported positive clinical outcome. We also found that patients' microbiomes often started off in a dysbiotic state, frequently characterized by the dominance of certain bacteria or viruses, but shifted to resemble healthy children as time went on. We also found that the microbiomes of these patients began to harbor higher levels of antibiotic resistance genes as time went on, up until around six months post-therapy. This likely correlates to high exposure levels of antibiotics administered prophylactically until the patients were no longer immunocompromised (around six months post-therapy). We also found high levels of adenovirus and astrovirus in these patients, which only occasionally corresponded to clinical reports. In each case, high viral loads subsided with time, indicating clearance of the virus.

This chapter demonstrated the potential applications for metagenomics in immune diseases and gene therapy. Because of established links between the microbiome and immune system, diseases that affect the immune system like SCID are likely to affect the immune system (and vice versa). In SCID patients, we were unsure whether immune system restoration would lead to restoration of a normal microbiome, or if a dysbiotic state would persist due to a founder-type effect. In the patients for which we had the longest sampling time course, we did see a normalization of the microbiome and decrease of antibiotic resistance load after immune reconstitution. While this conclusion is limited by the small number of patients, it suggests that the immune system acts to maintain a healthy microbial community in these patients.

Future directions

The work presented in this thesis center around the use of metagenomic sequencing in a variety of diseases. I explored the use of sequencing to robustly identify triggers in sarcoidosis, efficiently isolate difficult-to-culture pathogen genomes such as *Mycobacterium tuberculosis* using SWGA, and characterize microbe-immune interactions in SCID after gene therapy. These works center around short-read sequencing, the current state-of-the-art sequencing method used today for metagenomics. However, it seems likely that this will soon be supplanted by rising long-read technologies, such as those from Pacific Biosciences and Oxford Nanopore.

The essence of nanopore long-read sequencing is a small pore that moves nucleotide molecules through a few bases at a time and measures changes in the electrical potential based on the sequence. It avoids issues inherent to Illumina chemistry that lead to high error rates with longer sequences, and consequently can read tens of kilobases in length per sequence (under optimal conditions). However, nanopore sequencing comes with its own drawbacks, including a higher error rate and more difficult-to-model error profile. Even in the last few years, however, the error rates have decreased substantially and it seems reasonable to expect them to continue falling. Another key benefit of nanopore sequencing, specific to the Oxford minION, is the extremely small size of the sequencer itself. It is only slightly larger than a thumb drive and connects to a computer using USB. This small size raises alluring potential for field deployment as part of a portable sequencing and analysis toolkit.

Metagenomically, long-read sequencing offers a number of advancements over current methods. First, the retrieval of long contiguous DNA allows us to fill in the gaps and reconstruct microbial genomes much more easily (Loman, Quick, & Simpson, 2015; Quick et al., 2017). More comprehensive microbial genome databases will significantly improve our ability to classify and functionally annotate microbiomes. Second, long reads offer the ability to recover strain-level resolution in tagged sequencing experiments. While we are nominally constrained to tagged sequences smaller than 500bp in Illumina paired-end sequencing, long reads would enable the recovery of the entire 16S or ITS region and significantly improve our phylogenetic resolution (Kerkhof, Dillon, Häggblom, & McGuinness, 2017). Third, it opens up new opportunities for biosurveillance. Assuming we could identify a specific sequence unique to a pathogen or virulence gene of interest, long read sequencers can "read until" detection of that sequence (Loose, Malla, & Stout, 2016). Paired with a potentially field-deployable sequencer, this could enable extremely rapid monitoring of an environment for pathogens and antibiotic resistance capabilities, without the need for culture, extensive bioinformatics analysis, or data storage.

Even considering our current approaches in short-read sequencing, there remains significant room for advancement from the techniques outlined in this thesis. Metagenomic sequencing is exquisitely sensitive, but this can result in a higher level of false positives and spurious detections. In Chapter 2, I outlined an experimental approach that helps assuage these problems, but there are situations in which environmental controls may not be identifiable or obtainable. Approaches that restrict DNA to just those

from viable organisms may be worthwhile in certain studies (Emerson et al., 2017) but often samples are collected and stored in ways that damage or kill the microbes in them. Other ways of reducing spurious detections include the use of RNA in conjunction with DNA, on the premise that RNA is more readily degraded in the environment due to ubiquitous RNase activity. Making cDNA libraries from RNA and comparing to the results from gDNA could identify signatures from only authentically-present microbes in the original sample.

It is likely that metagenomic sequencing will become a core part of diagnostic and clinical work in time, but it has some way to go before it reaches the necessary level of affordability, speed and rigor. The promise of rapid, unbiased detection of any microbe or gene in a sample is one that could revolutionize our ability to track pathogens and disease outbreaks. Similarly, integration of metagenomic sequencing with new advancements in immunological sequencing will allow us to more effectively treat and understand immunological diseases. In total, the work demonstrated in this thesis lays a foundation for more robust and informative uses of metagenomic sequencing in the study of idiopathic and immunological disorders.

BIBLIOGRAPHY

Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors a unique microbiome. *Science translational medicine, 6*(237), 237ra265-237ra265. doi:10.1126/scitranslmed.3008599

Abbas, A. A., Diamond, J. M., Chehoud, C., Chang, B., Kotzin, J. J., Young, J. C., . . . Collman, R. G. (2016). The Perioperative Lung Transplant Virome: Torque Teno Viruses are Elevated in Donor Lungs and Show Divergent Dynamics In Primary Graft Dysfunction. *Am J Transplant*. doi:10.1111/ajt.14076

Aggarwala, V., Liang, G., & Bushman, F. D. (2017). Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob DNA, 8*, 12. doi:10.1186/s13100-017-0095-y

Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M. P., Baricordi, C., . . . Naldini, L. (2013). Lentiviral Hematopoietic Stem Cell Gene Therapy in Patients with Wiskott-Aldrich Syndrome. *Science*. doi:10.1126/science.1233151

Aiuti, A., Cassani, B., Andolfi, G., Mirolo, M., Biasco, L., Recchia, A., . . . Bordignon, C. (2007). Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J Clin Invest, 117*(8), 2233-2240. doi:10.1172/JCI31666

Aiuti, A., Slavin, S., Aker, M., Ficara, F., Deola, S., Mortellaro, A., . . . Bordignon, C. (2002). Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science, 296*(5577), 2410-2413. doi:10.1126/science.1070104

Allawi, H. T., & SantaLucia, J. (1997). Thermodynamics and NMR of internal GT mismatches in DNA. *Biochemistry, 36*(34), 10581-10594. doi:10.1021/bi962590c

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., . . . Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods, 11*(11), 1144-1146. doi:10.1038/nmeth.3103

Amann, R. I., Binder, B. J., Olson, R. J., Chisholm, S. W., Devereux, R., & Stahl, D. A. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and environmental microbiology, 56*(6), 1919-1925.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res, 9*(13), 3015-3027.

Arpaia, N., Campbell, C., Fan, X., Dikiy, S., van der Veeken, J., deRoos, P., . . . Rudensky, A. Y. (2013). Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature, 504*(7480), 451-455. doi:10.1038/nature12726

Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., . . . Honda, K. (2013). Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature, 500*(7461), 232-236. doi:10.1038/nature12331

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289-300.

Berry, C., Hannenhalli, S., Leipzig, J., & Bushman, F. D. (2006). Selection of target sites for mobile DNA integration in the human genome. *PLoS computational biology, 2*(11), e157.

Berry, C. C., Gillet, N. A., Melamed, A., Gormley, N., Bangham, C. R., & Bushman, F. D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics, 28*(6), 755-762. doi:bts004 [pii]

10.1093/bioinformatics/bts004

Berry, C. C., Nobles, C., Six, E., Wu, Y., Malani, N., Sherman, E., . . . Bushman, F. D. (2017). INSPIIRED: Quantification and Visualization Tools for Analyzing Integration Site Distributions. *Mol Ther Methods Clin Dev, 4*, 17-26. doi:10.1016/j.omtm.2016.11.003

Biffi, A., Montini, E., Lorioli, L., Cesani, M., Fumagalli, F., Plati, T., . . . Naldini, L. (2013). Lentiviral Hematopoietic Stem Cell Gene Therapy Benefits Metachromatic Leukodystrophy. *Science*. doi:10.1126/science.1233158

Bittinger, K., Charlson, E. S., Loy, E., Shirley, D. J., Haas, A. R., Laughlin, A., . . . Bushman, F. D. (2014). Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. *Genome Biol, 15*(10), 487. doi:10.1186/s13059-014-0487-y

Blevins, S. M., & Bronze, M. S. (2010). Robert Koch and the 'golden age' of bacteriology. *International Journal of Infectious Diseases, 14*(9), e744-e751. doi:10.1016/j.ijid.2009.12.003

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., . . . Fire, A. Z. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med, 1*(12), 12ra23.

Branton, W. G., Ellestad, K. K., Maingat, F., Wheatley, B. M., Rud, E., Warren, R. L., . . . Power, C. (2013). Brain Microbial Populations in HIV/AIDS: α-Proteobacteria Predominate Independent of Host Immune Status. *PloS one, 8*(1), e54673. doi:10.1371/journal.pone.0054673

Britton, R. A., & Young, V. B. (2014). Role of the intestinal microbiota in resistance to colonization by Clostridium difficile. *Gastroenterology, 146*(6), 1547-1553. doi:10.1053/j.gastro.2014.01.059

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., . . . Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature, 523*(7559), 208-211. doi:10.1038/nature14486

Buve, A., Jespers, V., Crucitti, T., & Fichorova, R. N. (2014). The vaginal microbiota and susceptibility to HIV. *AIDS, 28*(16), 2333-2344.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi:10.1186/1471-2105-10-421

Campos, P. F. G., T. (2011). DNA Extraction from Formalin Fixed Material *Ancient DNA* (Vol. 11, pp. 81-85): Humana Press.

Campregher, P. V., Srivastava, S. K., Deeg, H. J., Robins, H. S., & Warren, E. H. (2010). Abnormalities of the alphabeta T-cell receptor repertoire in advanced

myelodysplastic syndrome. *Exp Hematol, 38*(3), 202-212. doi:S0301-472X(09)00461-5 [pii]

10.1016/j.exphem.2009.12.004

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics, 26*(2), 266-267. doi:10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods, 7*(5), 335-336. doi:10.1038/nmeth.f.303

Cavazzana-Calvo, M., Andre-Schmutz, I., & Fischer, A. (2013). Haematopoietic stem cell transplantation for SCID patients: where do we stand? *British Journal of Haematology, 160*(2), 146-152. doi:10.1111/bjh.12119

Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., Yvon, E., Nusbaum, P., . . . Fischer, A. (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science, 288*(5466), 669-672.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics, 43*(4), 783-791.

Charlson, E. S., Bittinger, K., Haas, A. R., Fitzgerald, A. S., Frank, I., Yadav, A., . . . Collman, R. G. (2011). Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am J Respir Crit Care Med, 184*(8), 957-963. doi:10.1164/rccm.201104-0655OC

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., . . . Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS one, 5*(12), e15216. doi:10.1371/journal.pone.0015216

Charlson, E. S., Diamond, J. M., Bittinger, K., Fitzgerald, A. S., Yadav, A., Haas, A. R., . . . Collman, R. G. (2012). Lung-enriched organisms and aberrant bacterial and fungal respiratory microbiota after lung transplant. *Am J Respir Crit Care Med, 186*(6), 536-545. doi:10.1164/rccm.201204-0693OC

Chase, M. W. (1961). The preparation and standardization of Kveim testing antigen. *Am Rev Respir Dis, 84(5)Pt 2*, 86-88. doi:10.1164/arrd.1961.84.5P2.86

Chen, E. S., & Moller, D. R. (2014). Etiologic role of infectious agents. *Semin Respir Crit Care Med, 35*(3), 285-295. doi:10.1055/s-0034-1376859

Chen, E. S., & Moller, D. R. (2015). Etiologies of Sarcoidosis. *Clin Rev Allergy Immunol, 49*(1), 6-18. doi:10.1007/s12016-015-8481-z

Chen, E. S., Wahlstrom, J., Song, Z., Willett, M. H., Wiken, M., Yung, R. C., . . . Moller, D. R. (2008). T cell responses to mycobacterial catalase-peroxidase profile a pathogenic antigen in systemic sarcoidosis. *J Immunol, 181*(12), 8784-8796.

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., . . . Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics, 28*(16), 2106-2113. doi:10.1093/bioinformatics/bts342

Chiba, S., Okada, S., Suzuki, Y., Watanuki, Z., Mitsuishi, Y., Igusa, R., . . . Uchiyama, B. (2009). Cladosporium species-related hypersensitivity pneumonitis in household environments. *Intern Med, 48*(5), 363-367.

Chin, E. L., da Silva, C., & Hegde, M. (2013). Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. *BMC Genet, 14*, 6. doi:10.1186/1471-2156-14-6

Clarke, E. L., Lauder, A. P., Hofstaedter, C. E., Hwang, Y., Fitzgerald, A. S., Imai, I., . . . Collman, R. G. (2017). Microbial Lineages in Sarcoidosis: A Metagenomic Analysis Tailored for Low Microbial Content Samples. *Am J Respir Crit Care Med*. doi:10.1164/rccm.201705-0891OC

Cowell, A. N., Loy, D. E., Sundararaman, S. A., Valdivia, H., Fisch, K., Lescano, A. G., . . . Winzeler, E. A. (2017). Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed Clinical Samples. *mBio, 8*(1), e02257-02216. doi:10.1128/mBio.02257-16

Dean, F. B., Hosono, S., Fang, L. H., Wu, X. H., Faruqi, A. F., Bray-Ward, P., . . . Lasken, R. S. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America, 99*(8), 5261-5266. doi:10.1073/pnas.082089499

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., . . . Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology, 72*(7), 5069-5072.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science, 14*(6), 927-930. doi:DOI 10.1111/j.1654-1103.2003.tb02228.x

Dollive, S., Peterfreund, G. L., Sherrill-Mix, S., Bittinger, K., Sinha, R., Hoffmann, C., . . . Bushman, F. D. (2012). A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol, 13*(7), R60. doi:10.1186/gb-2012-13-7-r60

Douglas Bates, M. M., Ben Bolker and, & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software, 67*(1), 1—48. doi:10.18637/jss.v067.i01

Drake, W. P., Pei, Z., Pride, D. T., Collins, R. D., Cover, T. L., & Blaser, M. J. (2002). Molecular analysis of sarcoidosis tissues for mycobacterium species DNA. *Emerg Infect Dis, 8*(11), 1334-1341. doi:10.3201/eid0811.020318

Dubaniewicz, A. (2013). Microbial and human heat shock proteins as 'danger signals' in sarcoidosis. *Hum Immunol, 74*(12), 1550-1558. doi:10.1016/j.humimm.2013.08.275

Dubaniewicz, A., Trzonkowski, P., Dubaniewicz-Wybieralska, M., Dubaniewicz, A., Singh, M., & Mysliwski, A. (2007). Mycobacterial heat shock protein-induced blood T lymphocytes subsets and cytokine pattern: comparison of sarcoidosis with tuberculosis and healthy controls. *Respirology, 12*(3), 346-354. doi:10.1111/j.1440-1843.2007.01076.x

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461. doi:10.1093/bioinformatics/btq461

EL Clarke, M. R., KC Patterson, A Fitzgerald, M Feldman, L Litzky, F Bushman, RG Collman. (2015). Metagenomic analysis of Microbial Sequences in Specimens from Patients with Sarcoidosis and Controls. *Am J Respir Crit Care Med, 191*(A6163).

Emerson, J. B., Adams, R. I., Román, C. M. B., Brooks, B., Coil, D. A., Dahlhausen, K., . . . Rothschild, L. J. (2017). Schrödinger's microbes: Tools for distinguishing the

living from the dead in microbial ecosystems. *Microbiome, 5*(1), 86. doi:10.1186/s40168-017-0285-3
Ezike, D. N., Nnamani, C. V., Ogundipe, O. T., & Adekanmbi, O. H. (2016). Airborne pollen and fungal spores in Garki, Abuja (North-Central Nigeria). *Aerobiologia, 32*(4), 697-707. doi:10.1007/s10453-016-9443-5
Feng, H., Shuda, M., Chang, Y., & Moore, P. S. (2008). Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science, 319*(5866), 1096-1100. doi:10.1126/science.1152586
Feng, H., Taylor, J. L., Benos, P. V., Newton, R., Waddell, K., Lucas, S. B., . . . Moore, P. S. (2007). Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *Journal of virology, 81*(20), 11332-11340. doi:10.1128/JVI.00875-07
Fingerlin, T. E., Hamzeh, N., & Maier, L. A. (2015). Genetics of Sarcoidosis. *Clin Chest Med, 36*(4), 569-584. doi:10.1016/j.ccm.2015.08.002
Fischer, A., Grunewald, J., Spagnolo, P., Nebel, A., Schreiber, S., & Muller-Quernheim, J. (2014). Genetics of sarcoidosis. *Semin Respir Crit Care Med, 35*(3), 296-306. doi:10.1055/s-0034-1376860
Forde, B. M., & O'Toole, P. W. (2013). Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in Functional Genomics, 12*(5), 440-453. doi:10.1093/bfgp/els062
Fredricks, D. N., & Relman, D. A. (1996). Sequence-based identification of microbial pathogens: A reconsideration of Koch's postulates. *Clinical Microbiology Reviews, 9*(1), 18-&.
Gardes, M., & Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts. *Mol Ecol, 2*(2), 113-118.
Gaspar, H. B., Cooray, S., Gilmour, K. C., Parsley, K. L., Adams, S., Howe, S. J., . . . Thrasher, A. J. (2011). Long-term persistence of a polyclonal T cell repertoire after gene therapy for X-linked severe combined immunodeficiency. *Sci Transl Med, 3*(97), 97ra79. doi:3/97/97ra79 [pii]

10.1126/scitranslmed.3002715
Gaspar, H. B., Parsley, K. L., Howe, S., King, D., Gilmour, K. C., Sinclair, J., . . . Thrasher, A. J. (2004). Gene therapy of X-linked severe combined immunodeficiency by use of a pseudotyped gammaretroviral vector. *Lancet, 364*(9452), 2181-2187. doi:10.1016/S0140-6736(04)17590-9
Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., . . . Xavier, R. J. (2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe, 15*(3), 382-392. doi:10.1016/j.chom.2014.02.005
Ghazanfar, S., Azim, A., & Ghazanfar, M. A. (2010). Metagenomics and its application in soil microbial community studies: biotechnological prospects. *Journal of Animal & …, 6*(2), 611-622.
Gini, C. (1912). *Variabilità e mutuabilità. contributo allo studio delle distribuzioni e delle relazioni statistiche (variability and mutability. contribution to the study of the …*: Tipogr. di Cupini.
Gray, M. W., Sankoff, D., & Cedergren, R. J. (1984). On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved

structural core in small subunit ribosomal RNA. *Nucleic Acids Res, 12*(14), 5837-5852.
Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., . . . Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med, 7*, 99. doi:10.1186/s13073-015-0220-9
Guarner, F., & Malagelada, J. R. (2003). Gut flora in health and disease. *Lancet, 361*(9356), 512-519. doi:10.1016/S0140-6736(03)12489-0
Guggisberg, A. M., Sundararaman, S. A., Lanaspa, M., Moraleda, C., González, R., Mayor, A., . . . Odom, A. R. (2016). Whole-Genome Sequencing to Evaluate the Resistance Landscape Following Antimalarial Treatment Failure With Fosmidomycin-Clindamycin. *The Journal of infectious diseases, 214*(7), 1085-1091. doi:10.1093/infdis/jiw304
Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., . . . Schonrogge, K. (2015). PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods Ecol Evol, 6*(8), 973-980. doi:10.1111/2041-210X.12399
Haberman, Y., Tickle, T. L., Dexheimer, P. J., Kim, M.-O., Tang, D., Karns, R., . . . Denson, L. A. (2014). Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *The Journal of Clinical Investigation, 124*(8), 3617-3633. doi:10.1172/JCI75436
Hacein-Bey Abina, S., Gaspar, H. B., Blondeau, J., Caccavelli, L., Charrier, S., Buckland, K., . . . Cavazzana, M. (2015). Outcomes following gene therapy in patients with severe Wiskott-Aldrich syndrome. *JAMA, 313*(15), 1550-1563. doi:10.1001/jama.2015.3253
Hacein-Bey-Abina, S., Hauer, J., Lim, A., Picard, C., Wang, G. P., Berry, C. C., . . . Cavazzana-Calvo, M. (2010). Efficacy of gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med, 363*(4), 355-364. doi:10.1056/NEJMoa1000164
Hacein-Bey-Abina, S., Le Deist, F., Carlier, F., Bouneaud, C., Hue, C., De Villartay, J. P., . . . Cavazzana-Calvo, M. (2002). Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med, 346*(16), 1185-1193. doi:10.1056/NEJMoa012616
Hacein-Bey-Abina, S., Pai, S.-Y., Gaspar, H. B., Armant, M., Berry, C. C., Blanche, S., . . . Thrasher, A. J. (2014). A Modified γ-Retrovirus Vector for X-Linked Severe Combined Immunodeficiency. *N Engl J Med, 371*(15), 1407-1417.
Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., . . . Fischer, A. (2003). A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med, 348*(3), 255-256.
Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M. P., Wulffraat, N., Leboulch, P., . . . Cavazzana-Calvo, M. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science, 302*(5644), 415-419. doi:10.1126/science.1088547
Hannigan, G. D., Pulos, N., Grice, E. A., & Mehta, S. (2015). Current Concepts and Ongoing Research in the Prevention and Treatment of Open Fracture Infections. *Adv Wound Care (New Rochelle), 4*(1), 59-74. doi:10.1089/wound.2014.0531
Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ, 2*, e616. doi:10.7717/peerj.616

Hou, D., Zhou, X., Zhong, X., Settles, M. L., Herring, J., Wang, L., . . . Xu, C. (2013). Microbiota of the seminal fluid from healthy and infertile men. *Fertil Steril, 100*(5), 1261-1269. doi:10.1016/j.fertnstert.2013.07.1991

Howe, S. J., Mansour, M. R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., . . . Thrasher, A. J. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J Clin Invest, 118*(9), 3143-3150. doi:10.1172/JCI35798

Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., . . . Mazmanian, S. K. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell, 155*(7), 1451-1463. doi:10.1016/j.cell.2013.11.024

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., . . . Banfield, J. F. (2016). A new view of the tree of life. *Nat Microbiol, 1*, 16048. doi:10.1038/nmicrobiol.2016.48

Hume, J. C. C., Lyons, E. J., & Day, K. P. (2003). Human migration, mosquitoes and the evolution of Plasmodium falciparum. *Trends in parasitology, 19*(3), 144-149. doi:10.1016/S1471-4922(03)00008-4

Huttenhower, C., Knight, R., Brown, C. T., Caporaso, J. G., Clemente, J. C., Gevers, D., . . . White, O. (2014). Advancing the microbiome research community. *Cell, 159*(2), 227-230. doi:10.1016/j.cell.2014.09.022

Ishige, I., Usui, Y., Takemura, T., & Eishi, Y. (1999). Quantitative PCR of mycobacterial and propionibacterial DNA in lymph nodes of Japanese patients with sarcoidosis. *Lancet, 354*(9173), 120-123. doi:10.1016/S0140-6736(98)12310-3

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., . . . McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res, 45*(D1), D566-D573. doi:10.1093/nar/gkw1004

Josephson, K. L., Gerba, C. P., & Pepper, I. L. (1993). Polymerase chain reaction detection of nonviable bacterial pathogens. *Appl Environ Microbiol, 59*(10), 3513-3515.

Kalan, L., Loesche, M., Hodkinson, B. P., Heilmann, K., Ruthel, G., Gardner, S. E., & Grice, E. A. (2016). Redefining the Chronic-Wound Microbiome: Fungal Communities Are Prevalent, Dynamic, and Associated with Delayed Healing. *mBio, 7*(5). doi:10.1128/mBio.01058-16

Kamada, N., & Núñez, G. (2014). Regulation of the immune system by the resident intestinal bacteria. *Gastroenterology, 146*(6), 1477-1488. doi:10.1053/j.gastro.2014.01.060

Kaminski, J., Gibson, M. K., Franzosa, E. A., Segata, N., Dantas, G., & Huttenhower, C. (2015). High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol, 11*(12), e1004557. doi:10.1371/journal.pcbi.1004557

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ, 3*, e1165. doi:10.7717/peerj.1165

Kelly, B. J., Imai, I., Bittinger, K., Laughlin, A., Fuchs, B. D., Bushman, F. D., & Collman, R. G. (2016). Composition and dynamics of the respiratory tract microbiome in intubated patients. *Microbiome, 4*, 7. doi:10.1186/s40168-016-0151-8

Kennedy, M. K., Glaccum, M., Brown, S. N., Butz, E. A., Viney, J. L., Embers, M., . . . Peschon, J. J. (2000). Reversible defects in natural killer and memory CD8 T cell lineages in interleukin 15-deficient mice. *J Exp Med, 191*(5), 771-780.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), 996-1006. doi:10.1101/gr.229102

Kerkhof, L. J., Dillon, K. P., Häggblom, M. M., & McGuinness, L. R. (2017). Profiling bacterial communities by MinION sequencing of ribosomal operons. *Microbiome, 5*(1). doi:10.1186/s40168-017-0336-9

Kim, D., Hofstaedter, C. E., Zhao, C., Mattei, L., Tanes, C., Clarke, E., . . . Kelsen, J. (2017). Optimizing methods and dodging pitfalls in microbiome research. *Microbiome, 5*(1), 52.

Klein, J. T., Horn, T. D., Forman, J. D., Silver, R. F., Teirstein, A. S., & Moller, D. R. (1995). Selection of oligoclonal V beta-specific T cells in the intradermal response to Kveim-Siltzbach reagent in individuals with sarcoidosis. *J Immunol, 154*(3), 1450-1460.

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., . . . Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat Methods, 8*(9), 761-763. doi:10.1038/nmeth.1650

Koch, R. (1876). Investigations into bacteria: V. *The etiology of anthrax, based on the ontogenesis of Bacillus anthracis. Cohns Beitrage zur Biologie der Pflanzen, 2*(2), 277-310.

Koch, R., & Carter, K. C. (1987). *Essays of Robert Koch*. New York: Greenwood Press.

Koljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., . . . Ursing, B. M. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol, 166*(3), 1063-1068. doi:10.1111/j.1469-8137.2005.01376.x

Kovanen, P. E., & Leonard, W. J. (2004). Cytokines and immunodeficiency diseases: critical roles of the gamma(c)-dependent cytokines interleukins 2, 4, 7, 9, 15, and 21, and their signaling pathways. *Immunol Rev, 202*, 67-83. doi:10.1111/j.0105-2896.2004.00203.x

Kryazhimskiy, S., Rice, D. P., Jerison, E. R., & Desai, M. M. (2014). Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science, 344*(6191), 1519-1522. doi:10.1126/science.1250939

Landsteiner K, P. E. (1909). Uebertragung der Poliomyelitis acuta auf Affen. *Z Immunitätsforsch., 2*, 377-390.

Lauder, A. P., Roche, A. M., Sherrill-Mix, S., Bailey, A., Laughlin, A. L., Bittinger, K., . . . Bushman, F. D. (2016). Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome, 4*(1), 29. doi:10.1186/s40168-016-0172-3

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., . . . Pedersen, O. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature, 500*(7464), 541-546. doi:10.1038/nature12506

Leichty, A. R., & Brisson, D. (2014). Selective whole genome amplification for resequencing target microbial species from complex natural samples. *Genetics, 198*(2), 473-481. doi:10.1534/genetics.114.165498

Leidy, J. (1861). *A flora and fauna within living animals* (Vol. 44): Smithsonian institution.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., . . . Bushman, F. D. (2015). Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host & Microbe, 18*(4), 489-500. doi:10.1016/j.chom.2015.09.008

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi:10.1093/bioinformatics/btp698

Lodolce, J. P., Boone, D. L., Chai, S., Swain, R. E., Dassopoulos, T., Trettin, S., & Ma, A. (1998). IL-15 receptor maintains lymphoid homeostasis by supporting lymphocyte homing and proliferation. *Immunity, 9*(5), 669-676.

Loeffier, F., & Frosch, P. (1897). Summarischer Bericht Ober die Ergebnisse der Untersuchungen der Kommission zur Erforschung der Maul-und Klauenseuche bei dem Institut for Infektionskrankheiten in Berlin. *Zbl Bakt, 1*, 257-259.

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods, 12*(8), 733-735. doi:10.1038/nmeth.3444

Loose, M., Malla, S., & Stout, M. (2016). Real-time selective sequencing using nanopore technology. *Nat Methods, 13*(9), 751-754. doi:10.1038/nmeth.3930

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J, 5*(2), 169-172. doi:10.1038/ismej.2010.133

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics, 4*(12), 981-994. doi:10.1038/nrg1226

Makino, S., Chang, M. F., Shieh, C. K., Kamahora, T., Vannier, D. M., Govindarajan, S., & Lai, M. M. (1987). Molecular cloning and sequencing of a human hepatitis delta (delta) virus RNA. *Nature, 329*(6137), 343-346. doi:10.1038/329343a0

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *dx.doi.org, 9*(1), 387-402. doi:10.1146/annurev.genom.9.081307.164359

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal, 17*(1), pp. 10-12. doi:10.14806/ej.17.1.200

Martínez, F., Lafforgue, G., Morelli, M. J., González-Candelas, F., Chua, N.-H., Daròs, J.-A., & Elena, S. F. (2012). Ultradeep sequencing analysis of population dynamics of virus escape mutants in RNAi-mediated resistant plants. *Molecular Biology and Evolution, 29*(11), 3297-3307. doi:10.1093/molbev/mss135

Meisel, J. S., Hannigan, G. D., Tyldsley, A. S., SanMiguel, A. J., Hodkinson, B. P., Zheng, Q., & Grice, E. A. (2016). Skin Microbiome Surveys Are Strongly Influenced by Experimental Design. *J Invest Dermatol, 136*(5), 947-956. doi:10.1016/j.jid.2016.01.016

Michael Love, W. H., Simon Anders. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(550). doi:10.1186/s13059-014-0550-8

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D., & Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A, 110*(30), 12450-12455. doi:10.1073/pnas.1300833110

Mitchell, R. S., Beitzel, B. F., Schroder, A. R., Shinn, P., Chen, H., Berry, C. C., . . . Bushman, F. D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS biology, 2*(8), E234.

Mollerup, S., Friis-Nielsen, J., Vinner, L., Hansen, T. A., Richter, S. R., Fridholm, H., . . . Hansen, A. J. (2016). Propionibacterium acnes: Disease-Causing Agent or Common Contaminant? Detection in Diverse Patient Samples by Next-Generation Sequencing. *J Clin Microbiol, 54*(4), 980-987. doi:10.1128/JCM.02723-15

Naccache, S. N., Federman, S., Veeeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., . . . Chiu, C. Y. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research, 24*(7), 1180-1192. doi:10.1101/gr.171934.113

Nelson, M. I., Edelman, L., Spiro, D. J., Boyne, A. R., Bera, J., Halpin, R., . . . Holmes, E. C. (2008). Molecular Epidemiology of A/H3N2 and A/H1N1 Influenza Virus during a Single Epidemic Season in the United States. *Plos Pathogens, 4*(8), e1000133. doi:10.1371/journal.ppat.1000133

Nishiwaki, T., Yoneyama, H., Eishi, Y., Matsuo, N., Tatsumi, K., Kimura, H., . . . Matsushima, K. (2004). Indigenous pulmonary Propionibacterium acnes primes the host in the development of sarcoid-like pulmonary granulomatosis in mice. *Am J Pathol, 165*(2), 631-639. doi:10.1016/S0002-9440(10)63327-5

Niskanen, S., & Östergård, P. R. J. (2003). Cliquer User's Guide, Version 1.0. *users.aalto.fi.* Retrieved from http://users.aalto.fi/~pat/cliquer.html

Noguchi, M., Yi, H., Rosenblatt, H. M., Filipovich, A. H., Adelstein, S., Modi, W. S., . . . Leonard, W. J. (1993). Interleukin-2 receptor gamma chain mutation results in X-linked severe combined immunodeficiency in humans. *Cell, 73*(1), 147-157.

Nunes, M. R. T., Faria, N. R., Vasconcelos, H. B., Medeiros, D. B. d. A., Silva de Lima, C. P., Carvalho, V. L., . . . Vasconcelos, P. F. d. C. (2012). Phylogeography of dengue virus serotype 4, Brazil, 2010-2011. *Emerging Infectious Diseases, 18*(11), 1858-1864. doi:10.3201/eid1811.120217

Nureki, S., Miyazaki, E., Matsuno, O., Takenaka, R., Ando, M., Kumamoto, T., . . . Ezaki, T. (2007). Corynebacterium ulcerans infection of the lung mimicking the histology of Churg-Strauss syndrome. *Chest, 131*(4), 1237-1239. doi:10.1378/chest.06-2346

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res, 44*(D1), D733-745. doi:10.1093/nar/gkv1189

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. doi:10.1038/ismej.2017.126

Oyola, S. O., Ariani, C. V., Hamilton, W., Kekre, M., Amenga-Etego, L., Ghansah, A., . . . Kwiatkowski, D. P. (2016). Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *bioRxiv*, 067546. doi:10.1101/067546

Pammi, M., Cope, J., Tarr, P. I., Warner, B. B., Morrow, A. L., Mai, V., . . . Neu, J. (2017). Intestinal dysbiosis in preterm infants preceding necrotizing enterocolitis: a systematic review and meta-analysis. *Microbiome, 5*(1), 31. doi:10.1186/s40168-017-0248-8

Pannaraj, P. S., Li, F., Cerini, C., Bender, J. M., Yang, S., Rollie, A., . . . Aldrovandi, G. M. (2017). Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. *JAMA Pediatr, 171*(7), 647-654. doi:10.1001/jamapediatrics.2017.0378

Peternel, R., Culig, J., & Hrga, I. (2004). Atmospheric concentrations of Cladosporium spp. and Alternaria spp. spores in Zagreb (Croatia) and effects of some meteorological factors. *Ann Agric Environ Med, 11*(2), 303-307.

Ponstingl, H., & Ning, Z. (2010). SMALT [Utility]. Retrieved from http://www.sanger.ac.uk/science/tools/smalt-0

Puck, J. M., Deschênes, S. M., Porter, J. C., Dutra, A. S., Brown, C. J., Willard, H. F., & Henthorn, P. S. (1993). The interleukin-2 receptor gamma chain maps to Xq13.1 and is mutated in X-linked severe combined immunodeficiency, SCIDX1. *Hum Mol Genet, 2*(8), 1099-1104.

Puel, A., Ziegler, S. F., Buckley, R. H., & Leonard, W. J. (1998). Defective IL7R expression in T(-)B(+)NK(+) severe combined immunodeficiency. *Nat Genet, 20*(4), 394-397. doi:10.1038/3877

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., . . . Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc, 12*(6), 1261-1276. doi:10.1038/nprot.2017.066

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol, 35*(9), 833-844. doi:10.1038/nbt.3935

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Retrieved from https://www.r-project.org/

Relman, D. A., Schmidt, T. M., MacDermott, R. P., & Falkow, S. (1992). Identification of the uncultured bacillus of Whipple's disease. *The New England journal of medicine, 327*(5), 293-301. doi:10.1056/NEJM199207303270501

Renz, H., Brandtzaeg, P., & Hornef, M. (2011). The impact of perinatal immune development on mucosal homeostasis and chronic inflammation. *Nat Rev Immunol, 12*(1), 9-23. doi:10.1038/nri3112

Richter, E., Greinert, U., Kirsten, D., Rusch-Gerdes, S., Schluter, C., Duchrow, M., . . . Gerdes, J. (1996). Assessment of mycobacterial DNA in cells and tissues of mycobacterial and sarcoid lesions. *Am J Respir Crit Care Med, 153*(1), 375-380. doi:10.1164/ajrccm.153.1.8542146

Richter, E., Kataria, Y. P., Zissel, G., Homolka, J., Schlaak, M., & Muller-Quernheim, J. (1999). Analysis of the Kveim-Siltzbach test reagent for bacterial DNA. *Am J Respir Crit Care Med, 159*(6), 1981-1984.

Rieder, R., Wisniewski, P. J., Alderman, B. L., & Campbell, S. C. (2017). Microbes and mental health: A review. *Brain, behavior, and immunity, 66*, 9-17. doi:10.1016/j.bbi.2017.01.016

Rizk, G., Lavenier, D., & Chikhi, R. (2013). DSK: k-mer counting with very low memory usage. *Bioinformatics (Oxford, England), 29*(5), btt020-653. doi:10.1093/bioinformatics/btt020

Robins, H., Desmarais, C., Matthis, J., Livingston, R., Andriesen, J., Reijonen, H., . . . Cerosaletti, K. (2012). Ultra-sensitive detection of rare T cell clones. *J Immunol Methods, 375*(1-2), 14-19. doi:10.1016/j.jim.2011.09.001

Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., . . . Carlson, C. S. (2009). Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood, 114*(19), 4099-4107. doi:blood-2009-04-217604 [pii]

10.1182/blood-2009-04-217604

Robins, H. S., Srivastava, S. K., Campregher, P. V., Turtle, C. J., Andriesen, J., Riddell, S. R., . . . Warren, E. H. (2010). Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med, 2*(47), 47ra64. doi:2/47/47ra64 [pii]

10.1126/scitranslmed.3001442

Robinson, L. A., Smith, P., Sengupta, D. J., Prentice, J. L., & Sandin, R. L. (2013). Molecular analysis of sarcoidosis lymph nodes for microorganisms: a case-control study with clinical correlates. *BMJ Open, 3*(12), e004065. doi:10.1136/bmjopen-2013-004065

Rousseau, C., Poilane, I., De Pontual, L., Maherault, A.-C., Le Monnier, A., & Collignon, A. (2012). Clostridium difficile Carriage in Healthy Infants in the Community: A Potential Reservoir for Pathogenic Strains. *Clinical Infectious Diseases, 55*(9), 1209-1215. doi:10.1093/cid/cis637

Ruben J Colman, D. T. R. (2014). Fecal Microbiota Transplantation as Therapy for Inflammatory Bowel Disease: A Systematic Review and Meta-Analysis. *Journal of Crohn's & colitis, 8*(12), 1569. doi:10.1016/j.crohns.2014.08.006

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., . . . Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol, 12*, 87. doi:10.1186/s12915-014-0087-z

SanMiguel, A. J., Meisel, J. S., Horwinski, J., Zheng, Q., & Grice, E. A. (2017). Topical Antimicrobial Treatments Can Elicit Shifts to Resident Skin Bacterial Communities and Reduce Colonization by Staphylococcus aureus Competitors. *Antimicrob Agents Chemother, 61*(9). doi:10.1128/AAC.00774-17

Schmeisser, C., Steele, H., & Streit, W. R. (2007). Metagenomics, biotechnology with non-culturable microbes. *Applied Microbiology and Biotechnology, 75*(5), 955-962. doi:10.1007/s00253-007-0945-5

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., . . . Fungal Barcoding Consortium Author, L. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A, 109*(16), 6241-6246. doi:10.1073/pnas.1117018109

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., . . . Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature methods, 13*(5), 435-438. doi:10.1038/nmeth.3802

Schroder, A. R. W., Shinn, P., Chen, H. M., Berry, C., Ecker, J. R., & Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell, 110*(4), 521-529.

Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Mol Syst Biol, 9*, 666. doi:10.1038/msb.2013.22

Sellards, A. W., & Hindle, E. (1928). The Preservation of Yellow Fever Virus. *Br Med J, 1*(3512), 713-714.
Sherman, E., Nobles, C., Berry, C. C., Six, E., Wu, Y., Dryga, A., . . . Bushman, F. D. (2017). INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes. *Mol Ther Methods Clin Dev, 4*, 39-49. doi:10.1016/j.omtm.2016.11.002
Shreiner, A. B., Kao, J. Y., & Young, V. B. (2015). The gut microbiome in health and in disease. *Current opinion in gastroenterology, 31*(1), 69-75. doi:10.1097/MOG.0000000000000139
Siltzbach, L. E. (1961). The Kveim test in sarcoidosis. A study of 750 patients. *JAMA, 178*, 476-482.
Silva, C. L., & Ekizlerian, S. M. (1985). Granulomatous reactions induced by lipids extracted from Fonsecaea pedrosoi, Fonsecaea compactum, Cladosporium carrionii and Phialophora verrucosum. *J Gen Microbiol, 131*(1), 187-194. doi:10.1099/00221287-131-1-187
Smit, A. H., R; Green, P. (2013-2015). RepeatMasker Open-4.0.
Snow, J. (1856). The Mode of Propagation of Cholera. *The Lancet, 67*(1694). doi:10.1016/s0140-6736(02)67846-8
Song, Z., Marzilli, L., Greenlee, B. M., Chen, E. S., Silver, R. F., Askin, F. B., . . . Moller, D. R. (2005). Mycobacterial catalase-peroxidase is a tissue antigen and target of the adaptive immune response in systemic sarcoidosis. *J Exp Med, 201*(5), 755-767. doi:10.1084/jem.20040429
Stack, J. C., Murcia, P. R., Grenfell, B. T., Wood, J. L. N., & Holmes, E. C. (2012). Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proceedings of the Royal Society of London B: Biological Sciences, 280*(1750), rspb20122173-20122173. doi:10.1098/rspb.2012.2173
Suchankova, M., Paulovicova, E., Paulovicova, L., Majer, I., Tedlova, E., Novosadova, H., . . . Bucova, M. (2015). Increased antifungal antibodies in bronchoalveolar lavage fluid and serum in pulmonary sarcoidosis. *Scand J Immunol, 81*(4), 259-264. doi:10.1111/sji.12273
Sundararaman, S. A., Plenderleith, L. J., Liu, W., Loy, D. E., Learn, G. H., Li, Y., . . . Hahn, B. H. (2016). Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nature Communications, 7*. doi:10.1038/ncomms11078
Taylor, G. B., Paviour, S. D., Musaad, S., Jones, W. O., & Holland, D. J. (2003). A clinicopathological review of 34 cases of inflammatory breast disease showing an association between corynebacteria infection and granulomatous mastitis. *Pathology, 35*(2), 109-119.
Teirstein, A. S. (1998). Kveim antigen: what does it tell us about causation of sarcoidosis? *Semin Respir Infect, 13*(3), 206-211.
Tham, R., Katelaris, C. H., Vicendese, D., Dharmage, S. C., Lowe, A. J., Bowatte, G., . . . Erbas, B. (2017). The role of outdoor fungi on asthma hospital admissions in children and adolescents: A 5-year time stratified case-crossover analysis. *Environ Res, 154*, 42-49. doi:10.1016/j.envres.2016.12.016
The Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. *486*, 207. doi:10.1038/nature11234

https://www.nature.com/articles/nature11234 - supplementary-information

Thrasher, A. J., Gaspar, H. B., Baum, C., Modlich, U., Schambach, A., Candotti, F., . . . Fischer, A. (2006). Gene therapy: X-SCID transgene leukaemogenicity. *Nature, 443*(7109), E5-6; discussion E6-7. doi:10.1038/nature05219

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research, 27*(4), 626-638. doi:10.1101/gr.216242.116

Underhill, D. M., & Iliev, I. D. (2014). The mycobiota: interactions between commensal fungi and the host immune system. *Nature Reviews Immunology, 14*(6), 405-416. doi:10.1038/nri3684

Walker, W. A. (2013). Initial intestinal colonization in the human infant and immune homeostasis. *Ann Nutr Metab, 63 Suppl 2*, 8-15. doi:10.1159/000354907

Wang, G. P., Berry, C. C., Malani, N., Leboulch, P., Fischer, A., Hacein-Bey-Abina, S., . . . Bushman, F. D. (2010). Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood, 115*(22), 4356-4366. doi:10.1182/blood-2009-12-257352

Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C., & Bushman, F. D. (2007). HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research, 17*(8), 1186-1194.

Wang, G. P., Garrigue, A., Ciuffi, A., Ronen, K., Leipzig, J., Berry, C., . . . Bushman, F. D. (2008). DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic acids research, 36*(9), e49.

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., . . . Hazen, S. L. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature, 472*(7341), 57-63. doi:10.1038/nature09922

Weinstein, J. A., Jiang, N., White, R. A., 3rd, Fisher, D. S., & Quake, S. R. (2009). High-throughput sequencing of the zebrafish antibody repertoire. *Science, 324*(5928), 807-810. doi:324/5928/807 [pii]

10.1126/science.1170020

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol, 173*(2), 697-703.

Wilson, M. R., Naccache, S. N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., . . . Chiu, C. Y. (2014). Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. *proxy.library.upenn.edu, 370*(25), 2408-2417. doi:10.1056/NEJMoa1401268

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol, 15*(3), R46. doi:10.1186/gb-2014-15-3-r46

Young, J. C., Chehoud, C., Bittinger, K., Bailey, A., Diamond, J. M., Cantu, E., . . . Collman, R. G. (2015). Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am J Transplant, 15*(1), 200-209. doi:10.1111/ajt.13031