Expanding the Limits of Scale and Sensitivity in Microbial Genomics

by

Georgia Kerasia Lagoudas

B.S. Bioengineering
Rice University 2012

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Signature of Author...
Department of Biological Engineering
May 18, 2018

Signature redacted

Certified by...
Paul C. Blainey
Associate Professor of Biological Engineering
Thesis Supervisor

Signature redacted

Accepted by...
Forest White
Professor of Biological Engineering
Chair of Graduate Program, Department of Biological Engineering

Expanding the Limits of Scale and Sensitivity in Microbial Genomics

by

Georgia Kerasia Lagoudas

Submitted to the Department of Biological Engineering
on May 18, 2018 in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

Sequencing of microbial genomes has enabled new understanding of human health and disease. Certain microbes can support human health through the microbiota, helping to train our immune system or supply essential nutrients. In other cases, microbes may be pathogenic, overwhelming the immune system and causing infection. Low-cost and accessible DNA sequencing has allowed us to learn important information about microbial systems – we can identify what microbes are members of our microbiota and how they change with disease, as well as how pathogenic microbes evolve and acquire resistance to antibiotics. While the cost of sequencing has decreased and allowed for widespread use, studies are now limited by sample acquisition and preparation. In particular, microbial sample preparation has challenges at the limits of sensitivity (low signal to noise ratio) and at the limits of scale (large sample size). In this thesis, I developed methods to address both of these challenges and applied the techniques to study questions in basic biology and in clinical medicine. First, I developed a procedure to sample and sequence the lung microbiome in mouse models, where high background of mammalian DNA in lung samples poses a serious challenge for sequencing preparation. Along with my collaborator, I used this procedure to investigate the microbiome in a murine model of lung cancer. Second, I developed a platform for high-throughput sequencing preparation of bacteria at the scale of thousands of samples, with a 100-fold less cost per sample. I prepared and sequenced 3000 antibiotic-resistance bacteria from a clinical trial studying the role of decolonization procedures. This work provides new insights about microbes in the context of health and disease, and the methods developed here can make samples newly accessible for sequencing at the limits of scale or sensitivity.

Thesis Supervisor: Paul C. Blainey

Title: Associate Professor of Biological Engineering

Acknowledgments

First, thank you to my thesis advisor, Professor Paul Blainey. He provided a lab environment and scientific environment that was open, supportive, and welcoming to new ideas. Thank you to the Blainey Lab members – I am grateful for the scientific advice, challenging questions, and constant support. In particular, thank you to Dr. Lily Xu for her unwavering support and friendship – she brought life and energy to everything. Special thanks to the early Blainey Lab members, Nav Ranu, Tony Kulesa, David Feldman, Dr. Soohong Kim, Jacob Borrajo, and Francis McCarthy, for supporting me over these many years and making our lab a fun place to work, grow, and play. I was very fortunate to work with eight excellent undergraduates during my PhD – their work has helped shape my research. Special thanks to Rebecca Noel, Aman Patel, Nova Xu, Uriel Sanchez, and Andrea Li – you may not have known, but I learned as much from you as you did from me.

Thank you to my committee members, Profs. Eric Alm and Wendy Garrett, for their support during my PhD. Their advice helped shaped my research for the better, and I appreciate their positive energy and true thoughtfulness. Thank you to my collaborators at Harvard, Dr. Mohamad Sater and Prof. Yonatan Grad, for providing a unique opportunity to work together on clinical genomics. Thank you to my collaborators at the Koch Institute, Dr. Chengcheng Jin and Prof. Tyler Jacks, for fostering an exciting collaboration to study the lung microbiome.

I would like to thank the MIT Biological Engineering Department for fostering a positive, healthy environment. Profs. Doug Lauffenburger, Forest White, and Mark Bathe have made a true impact. Thank you to the BE Communication Lab, especially Jaime Goldstein, for shaping me early in my career. I am fortunate to have been part of an amazing team with the BE REFS, from which I will take away life-long skills. Thank you to Claire Duvallet for always being there 110%. I am grateful to Libby Mahaffy for shaping my perspective on life, relationships, and communication. Thank you to the BE class of 2012 for making us feel like family, with life-long friendships. Special thanks to Marianna Sofman for her positive energy, kind words, and adventurous spirit. Many friends have supported me along the way – thank you to Elyse Landry, Evi Van Itallie, Katie Mass, and Jen Wilson.

I would like to thank my mentors from Rice University – Kate Abad, Dean Hutchinson, Prof. Matteo Pasquali, Prof. Tony Mikos, and Prof. Junichiro Kono. And thank you to my early scientific mentors, Profs. Richard Behringer, Shigeo Maruyama, Luke Lee, and Robert Langer.

Importantly, I am grateful for the unwavering support from Scott Olesen over the past four years. Little did I know that I would find a life-long partner here at MIT, and his intellect, thoughtfulness, and wit have inspired me. And thank you to my brother-in-law, Prof. Justin Wilkerson, for his support and wise words.

Above all, thank you to my family. My sister, Natasha, and my parents, Magdalini and Dimitris, have given me tremendous support. They inspire me every day. They provided love, energy, and motivation and encouraged me to shoot for the stars. And thank you to my namesake, Γιώργος Ζαμπούκης, for doing cartwheels in the field. That has brought me here.

Thank you to the funding from the MIT Presidential Fellowship, the MIT Hugh Hampton Young Fellowship, and the NSF Graduate Research Fellowship that helped support this work.

Contents

Chapter 1 Introduction

A vast array of microbes inhabits the earth, serving a critical role in our ecosystem and in human health. Microbes, consisting of bacteria and archaea, are single-celled organisms that cover every corner of our world, from the cold, harsh environment of the deep sea to the nutrient-rich microcosm of the human intestine. Scientists have aimed to characterize and understand these microbes for hundreds of years, ultimately learning how to manipulate them to prevent or cure human disease. Humanity has discovered that microbes can be devastatingly destructive or critically essential in human health.

Microbes have a symbiotic relationship with humans, living on or within humans for both good and bad. The relationship can vary from mutualistic (both host and microbe benefit) to commensalistic (only microbe benefits) to parasitic (microbe benefits at a cost to the host). In a parasitic relationship, microbial pathogens have the capacity to infect humans, evading the immune system and potentially causing death. We have a stark example with the plague (caused by the bacterium *Yersinia pestis*) that caused three major pandemics over the past two centuries, killing tens of millions of people [1]. Other common modern-day infections may be caused by bacteria such as *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Escherichia coli,* and *Clostridium difficile*. On the other hand, microbes do have a positive and necessary role in human health, particularly through the microbiota, the collection of microbes that live on and within us. The microbiota aids us in digesting nutrients to make them bioavailable, trains our immune system to detect bacterial friend from foe, and plays a role in sustaining the healthy ecological niche that might otherwise be invaded by pathogenic bacteria. Wiping out this commensal microbiota can have clear implications in health, such as increasing susceptibility to asthma or to antibiotic-resistant *C. difficile* infections [2]–[6].

Studying these microbes, whether to find ways to promote their healthy growth or inhibit harmful growth, is a critical field of research. For example, the discovery of antibiotics, a powerful tool against pathogens, has saved millions from bacterial infections [7], [8]. However, bacteria evolve at a rapid rate and have developed mechanisms to evade cell death, often enabling swift acquisition of resistance to newly introduced antibiotics. This development of antibiotic resistance is a key question in human health – how can we predict, track, and respond to evolution of antibiotic resistance? We can study microbes to examine how their DNA changes, or mutates, over time and from learning how they evolve resistance, use this knowledge to better inform the development of future antibiotics.

1.1 Genomics can provide powerful insight to microbiology

DNA sequencing has powered a new biological revolution. Before the 1970s, we were limited to examining cells through molecular biology assays or phenotypic assays – measuring growth under certain conditions, testing for production of specific compounds, searching for the presence of specific proteins, or examining gel patterns for cut sites of DNA, for example. However, with the advent of DNA sequencing around 1976 [9], [10], we could move beyond information that was available only through phenotypic assays and instead read the entire sequence of letters (nucleotides) that encodes for cellular function. This sequence of nucleotides, the genome, provides the code for cells to perform all functions, whether it is making proteins that transport nutrients into cells or sending signals for cell growth or death.

Before DNA sequencing, microbial detection was limited by the need to culture microbes in order to analyze and characterize them. However, the vast majority of microbes in our environment have yet to be cultured (up to 99%) [11]. With DNA sequencing, we are now able to detect microbes without the necessity of culture. Additionally, sequencing allows us to investigate new types of systems – we can examine mixed populations of bacteria, dead bacteria, or DNA outside of cells. Sequencing uniquely positions biologists to study DNA in nearly any system, without the requirement of needing living, culturable cells. Furthermore, sequencing makes it possible to study very low quantities of DNA through amplification cycles followed by sequencing. For example, one group performed single-cell sequencing on microbes from nine diverse habitats, including a hot spring and a hydrothermal vent, that had not been successfully cultured. The single-cell sequencing identified 18 new candidate phyla on the microbial tree of life and uncovered a novel stop codon assignment [12]. Additionally, sequencing of unculturable bacteria in a marine sponge uncovered a new bacterial group that is a super-producer of bioactive compounds and provides opportunities for drug discovery [13].

The advent of DNA sequencing powered the Human Genome Project in the 1990s, a momentous 15-year effort to sequence the human genome. In recent years, the Human Microbiome Project (HMP), funded by the U.S. National Institutes of Health in 2009, aimed to sequence the variety of microbes found in and on humans [14]. The human microbiome is a collection of roughly 10 trillion microbes that live alongside a healthy person in places like the intestine, skin, mouth, vagina, and lung [15], [16]. The HMP collected samples from 240 patients across 15 different body sites, generating over 3.5 terabases of genomic data [16]. This dataset has opened the doors for researchers to explore and characterize the microbiome, whether searching for principled explanations of dynamics over time or new insights for therapeutics. For example, analysis of the collective microbial genomes identified thousands of biosynthetic gene clusters, including some with the ability to produce antibiotics, and lactocillin was identified as a novel antibiotic [17].

1.2 DNA sequencing is now widely accessible

Sequencing is ubiquitous and becoming a standard method in the biological toolkit for science research. Since the development of massively parallel or "next generation" sequencing (NGS) in 2005 [18], [19], the cost of sequencing has plummeted by five orders of magnitude (Figure 1-1)[20]. A megabase of DNA (one million basepairs) now costs roughly $0.01 to generate, resulting in a cost of about $2 per microbial genome (at an average genome coverage of 50x). As sequencing equipment is spreading to more research facilities and service providers, NGS is widely accessible both in terms of cost and equipment. There are established protocols for sequencing, and current Illumina sequencing systems offer a streamlined process that only requires the user to input a set of reagents and the DNA "library" to be sequenced, without any technical steps from the user. This entails that a graduate student can easily get trained on and use a sequencing machine to generate terabytes of data in mere days, an order of magnitude more data than the entire Human Genome Project.

Sequencing cost continues to drop while data generation capacity continues to grow, and we have not yet exhausted ideas for new technologies to increase scale and decrease cost. The newest technology platform from Illumina can generate more than a terabyte of data in a matter of days for a few thousand dollars. In the context of microbial research, this translates to over five thousand microbial genomes (at an average coverage of 50). Most studies do not even have this many bacterial samples to study, showing that the "sky is the limit" for sequencing microbial genomes. This is in contrast to human genomics research, where the size of the human genome (1000-fold larger than the average microbial genome) makes sequencing cost a limiting factor.

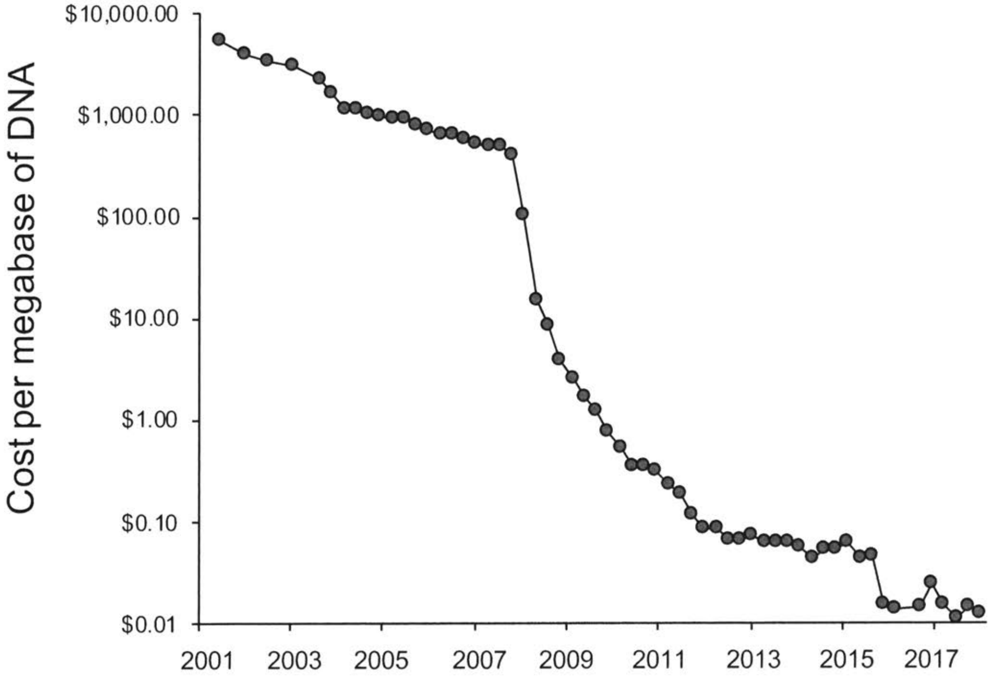


Figure 1-1. Sequencing cost per megabase of DNA (1 million basepairs) over time. In 2008, sequencing centers began transitioning to next generation sequencing technologies. Data from the National Human Genome Research Institute [20].

1.3 Preparing DNA for sequencing is the new bottleneck

Although sequencing technology is established, easy-to-use, and low-cost for microbial genomes, a DNA "library" is required for NGS. Preparing this DNA library is the new bottleneck, whereas previously the cost and throughput of sequencing was the major hurdle. A DNA library is a collection of DNA molecules that are of appropriate length and have adapters at each end necessary for Illumina sequencing. In this thesis, the focus will only be on Illumina sequencing technology. Although there are other sequencing platforms such as PacBio, Illumina currently offers the highest throughput and lowest per-base cost and as a result is the dominant NGS platform.

NGS allows for either sequencing one gene target (amplicon) or shotgun sequencing, where all the DNA from a mixture is sequenced together (Figure 1-2). A type of shotgun sequencing is whole genome sequencing, where all the genetic material from one cell type is sequenced. I will focus on whole genome sequencing (WGS) in this work, but other types of shotgun sequencing include metagenomics, where all genetic material from a mixture of organisms is sequenced together. In amplicon sequencing, it is common to use a selective amplification step to make many copies of the target of interest from the DNA sample. In whole genome sequencing, the entire genetic material from a cell is amplified together and cut up into similar size strands.

To provide an example of when targeted sequencing or whole genome sequencing might be used, consider the case of human genomics in the clinic. Targeted gene sequencing may be performed on tumor samples from a cancer patient to identify any specific mutations that might exist in a known set of cancer-related genes. Targeted sequencing requires less DNA to be sequenced, so it is less expensive and preferable when it can answer the question at hand. In contrast, if clinicians do not know what gene targets to look for, whole genome sequencing may be used to scan for mutations across the entire genome. WGS may be used in pediatric patients with a phenotype suspected to be caused by a Mendelian disorder, a single gene disorder that may inherited.

In order to generate a DNA library, a researcher much lyse the cells, extract and purify the DNA, selectively amplify target DNA (for amplicon sequencing) or randomly fragment/tag/amplify all DNA (for whole genome sequencing), followed by DNA purification, barcoding, quantification, and pooling with other samples. This long list of steps is in contrast to actually performing the sequencing: denature DNA, dilute to the proper concentration, and load onto a sequencing machine. In amplicon sequencing, challenges may arise when selectively amplifying the target DNA; for example, background DNA may be present that amplifies along with the target or generates a background noise that "swamps" the signal. In whole genome sequencing, the challenge arises due to the specialized enzymes required for fragmenting and tagging the DNA and costs associated with this procedure.

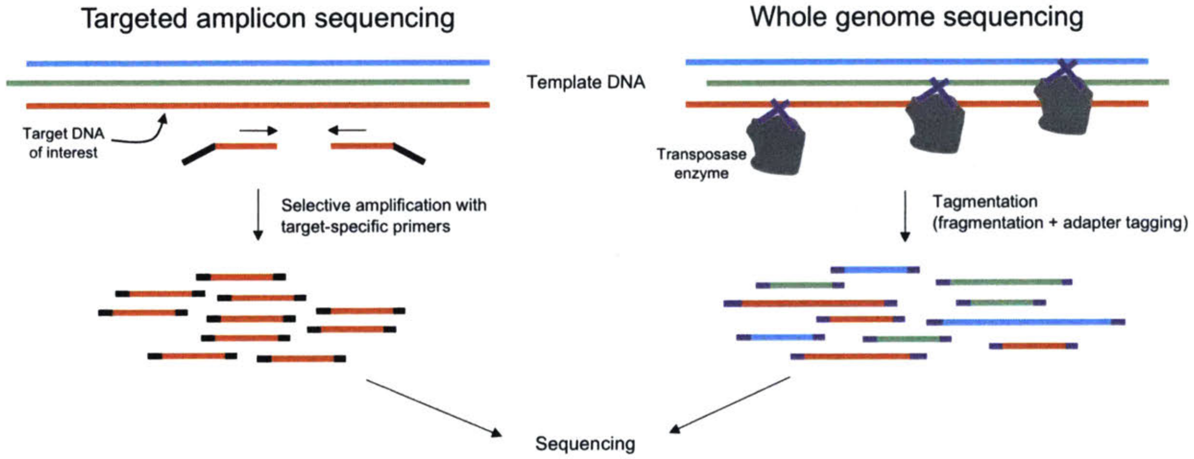

Targeted amplicon sequencing
Whole genome sequencing
Template DNA
Target DNA of interest
Transposase enzyme
Selective amplification with target-specific primers
Tagmentation (fragmentation + adapter tagging)
Sequencing

Figure 1-2. DNA sequencing preparation for targeted amplicon or whole genome sequencing.

While amplicon sequencing requires common polymerase enzymes and primers, whole genome sequencing requires specialized enzymes. The Nextera technology (developed by Illumina) includes an enzyme that simultaneously fragments the DNA and tags the DNA with sequencing adapters. This "tagmentation" reaction uses a transposase enzyme that randomly cuts DNA and inserts a transposon sequence at the end of the molecule (Figure 1-2) [21]. This technology allows for minimal bias and relatively even sampling of DNA molecules across the genome. Since both steps are included in one reaction, the process is more streamlined compared to previous methods and has revolutionized the ease of WGS preparation across many sample types. However, while the Nextera technology is effective, the cost of the enzyme and reagents are high and many steps are still required. Overall, the cost of preparing a sample for WGS starting from cells requires $100-$300 per sample (including reagents and labor). At large batch sizes this cost can drop slightly, but the Nextera reagent costs and labor still keep the price high. This is in contrast to the cost of actually sequencing a bacterial sample, which is two orders of magnitude less (about $2) (Figure 1-3). While there are new sequencing platforms currently under development that do not require any sample preparation besides DNA extraction and purification, these are still in their infant stage [22].

Therefore, both amplicon sequencing and WGS have a sample preparation bottleneck. While the cost of sequencing a microbial sample is low and sequencing itself is primarily automated, the new hurdle is sample preparation. In amplicon sequencing, accessing and amplifying the DNA of interest can be problematic, and for WGS the library preparation process is laborious and expensive for large sample sizes

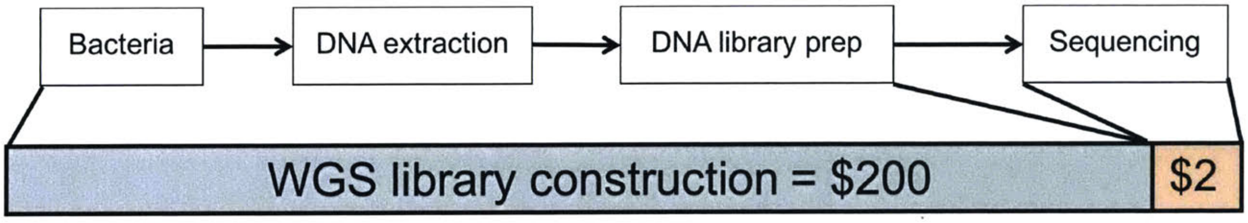


Figure 1-3. DNA library construction is the new bottleneck for sequencing.

1.4 Sensitivity or scale pose serious challenges

As described above, microbial WGS and amplicon sequencing are no longer limited by sequencing technology or cost, but rather by the library construction that precedes it. Overall, there are three general categories that may pose serious challenges with sequencing preparation:

1) **Accessing the DNA**
    a. Hard-to-lyse cells
    b. Low starting amount of DNA – samples that come from unculturable bacteria or slow to grow bacteria
    c. Low signal to noise ratio for DNA of interest – the DNA of interest may be mixed with other DNA (such as a small number of bacteria found in a human tissue sample)
2) **Large sample size** – for greater than 10s or 100s of samples, DNA library preparation cost and throughput is a serious hurdle
3) **Rapid turn-around time**

In this thesis, I address challenges at the limits of sensitivity (low signal to noise ratio) and at the limits of scale (large sample size). To provide a larger context for this work, I will briefly discuss the other challenges below.

Some bacteria that have outer cell walls that are very resistant to lysis, such as *Bacillus* and *Mycobacterium* [23]–[25]. Generally, multiple methods of lysis may be needed, such as chemical, enzymatic, or mechanical lysis. In the case that cells will not lyse with chemical or enzymatic methods, physical disruption of the membrane is generally sufficient. This is commonly achieved through bead beating, a procedure where small beads are added to a liquid suspension of cells and shaken vigorously. Bead beating is challenging for small sample volumes or with high-throughput processing, putting a limit of the number of samples that can be feasibly processed per unit time or cost.

Additionally, low amounts of starting DNA can pose a challenge to amplicon and whole genome sequencing. With low DNA quantities, many cycles of amplification are required to make enough copies of the target DNA for sequencing, potentially introducing bias. Furthermore, the library preparation process is inefficient, as there are multiple sequential steps that only carry over a portion of the DNA. When the starting quantity is low, losing half or more of the initial material can change the sequencing readout and distort the true signal. In addition, WGS library preparation with Nextera requires a minimum amount of starting DNA to have an appropriate ratio of DNA to enzyme. With low DNA quantities, the reaction volume becomes very small and it is difficult to pipet such low quantities without specialized robotics or microfluidics. Bacteria isolated from soil are often slow-growing or uncultivatable but provide opportunities for antibiotic discovery [26], [27], so sequencing preparation is important but challenging. Single cell sequencing is a subtopic of this category but will not be discussed here.

The third category, rapid turn-around time for sequencing, is also an important challenge. There is a need, especially in the clinic and in the field, for small batch size rapid sequencing (generally, in the range of 1-10 samples). New technologies, such as nanopore sequencing, are being developed that may provide on-demand sequencing with minimal sample preparation. The continued development of technologies to provide rapid sequencing results will have practical applications in medicine, but further discussion is outside the scope of this work.

While the challenges described above are important, this thesis will focus on addressing challenges at the limits of sensitivity (low signal to noise ratio) and at the limits of scale (large sample size). In my thesis, I developed methods in both of these areas to address questions in basic biology and in clinical medicine. First, I developed a procedure to sample and sequence the lung microbiome in mouse models, where high background of mammalian DNA in lung samples pose a serious challenge for sequencing preparation. Second, I developed a platform for high-throughput sequencing of bacteria at the scale of thousands of samples, with a 100-fold less cost per sample.

1.5 At the limits of sensitivity: sequencing preparation of lung microbiome samples

With a low signal to noise ratio of the DNA of interest, the target DNA is outnumbered by other non-target DNA that pose a library construction challenge. For example, identifying the bacteria in infected human tissue samples, examining microbes found within the tumor tissue microenvironment, or collecting lung or skin samples for microbiome analysis are all cases where the target of interest (bacteria) is mixed with a high quantity of host (human) DNA [28]–[32]. While these samples are challenging to work with, sequencing the microbial DNA is informative and can provide actionable clinical information for antibiotic treatment of infected tissues, for example. A specific case is with *Mycobacterium tuberculosis* – not only is the bacterium hard to lyse, but the sputum samples collected in the clinic harbor low quantities of bacteria and high amounts of host DNA. Researchers have raised a call to action for WGS sample preparation methods to detect mutations associated with drug resistance [25].

Potential approaches to this sequencing challenge include: 1) sequence everything in the system, at very high depth; 2) perform selective amplification of the target DNA for amplicon sequencing; or 3) deplete the background noise (host DNA). In the first case, sequencing deeply is only feasible when the ratio of signal to noise is high enough (on the order of 1:10) such that sequencing costs do not balloon out of proportion. When background DNA far outnumbers the target DNA, then deep sequencing is not economically feasible. In the second case, selective amplification can be very effective, especially since microbial DNA has many regions unique from mammalian DNA. However, this approach limits one to amplicon sequencing only and additional problems can arise if selective amplification is challenging to achieve. In the third case, depletion of background (mammalian) DNA may be appropriate but can also cause loss of microbial DNA unintentionally. Therefore, when investigating a new system this method is not ideal because loss of microbial DNA can distort the original signal.

In the second chapter of this thesis, I present methods for sensitive detection of microbial DNA from lung samples in mice. The lung microbiome is the collection of microbes that reside in the healthy lung.

Historically, the healthy lung was considered sterile because culture-based methods failed to detect bacteria [33]. However, with the application of DNA sequencing, bacteria have been detected in the healthy lung and are now widely acknowledged as common residents of both the upper and lower respiratory tract [34]. The lung microbiome has been linked to the proper development of the immune system, where absence of healthy microbe exposure in the lungs can lead to asthma [35], as well as to proper alveolar structure development [36]. Furthermore, the lung microbiota has been associated with altered disease states in chronic obstructive pulmonary disease and cystic fibrosis [30], [37]–[40]. The main issue that has delayed research and slowed progress in this field has been the very low bacterial biomass in the lung, especially compared to the gut and skin.

Due to the low quantity of bacteria in the lung, lung tissue samples or bronchoscopy samples have a far greater number of mammalian cells compared to microbial cells. This creates a challenge in sample preparation for sequencing, as there is a low signal to noise ratio (on the order of 1 bacterial DNA molecule per 10,000 mammalian DNA molecules). To overcome this challenge, researchers in the field use targeted DNA amplification of the 16S rRNA bacterial gene. However, many rounds of amplification are necessary and high amounts of contaminant DNA can be unintentionally amplified. Therefore, there is a need for systematic optimization of how to collect the lung sample, what sample type to use, what selective amplification conditions to use, and how to process sequencing data to account for any contaminants.

In Chapter 2 of this thesis, I describe methods developed for high sensitivity sequencing detection of microbial DNA in the mouse lung. I apply this optimized protocol to examine the lung microbiota in immunodeficient mouse models and determine whether an altered immune system causes a change in the lung microbiota. In Chapter 3 of this thesis, I subsequently apply these lung microbiome methods to examine the microbiota in the context of a specific lung disease, lung cancer. In collaboration with the lab of Tyler Jacks, I investigate the microbiota in a mouse model of lung cancer and find that the microbial quantity and diversity is altered in the diseased state.

1.6 At the limits of scale: high-throughput sequencing preparation of thousands of microbial samples

The fourth category of sample preparation challenge arises with large-scale sequencing. With increasing numbers of bacterial samples, cost and time become prohibitive for sequencing preparation. As outlined above, sample preparation of bacterial samples for sequencing is 100-fold more expensive than sequencing itself. However, large sample size sequencing is desirable in many cases, especially in the clinic. For instance, large-scale studies (>100 bacterial samples) have powered the earlier detection of pathogen outbreaks [41], accurate characterization of transmission networks [42], [43], detection of the timing and mechanism of antibiotic resistance acquisition [44], and bacterial genome-wide association studies [45]–[47]. Accessible large-scale sequencing in the clinic can improve infectious disease diagnostics and treatment, as well as enable molecular epidemiology studies.

Bacterial isolates (samples of one specific bacteria) are commonly collected in the clinic to evaluate the type of pathogen a patient carries. Isolates are collected by culturing, or growing, one specific bacteria from a patient sample, commonly on an agar plate or in a tube with liquid broth. These isolates of pure bacteria (from common pathogens like *E. coli, S. aureus,* and *P. aeruginosa)* are one the easiest types of samples to prepare for sequencing – they have a large amount of starting material, they can be recultured for additional material as needed, and due to their common presence as pathogens there are established protocols for lysing and extracting DNA. The Nextera protocol can be easily used to prepare a sample for WGS. However, scaling up the sample preparation has the challenges described above and is cost-prohibitive for many laboratories and clinics. In order to make sequencing accessible for large-scale studies, methods are needed for efficient preparation of DNA sequencing libraries from bacteria.

In Chapter 4 of this thesis, I describe the development of a microfluidic platform for whole genome sequencing library preparation. Through reduced reaction volume, integrated sample preparation, and automation, thousands of bacterial genomes can be prepared for sequencing in a few weeks. The microfluidic sample preparation enabled nearly 100-fold reduction in cost per sample, processing at a rate of 1000 samples per 7 days. We were motivated by the clinical application of using WGS to study pathogen transmission and recolonization dynamics in patients. I applied this technology to sequence 3,000 methicillin-resistant *Staphylococcus aureus* (MRSA) samples from a clinical trial studying the impact of decolonization protocols. Whole genome analysis was used to determine specific strain types and antibiotic resistance profiles. Our technology provides an inexpensive and reliable method for whole genome sequencing of bacterial isolates to answer new questions in clinical medicine.

Chapter 2

Methods for mouse lung microbiome characterization and demonstration of altered diversity in immunodeficient mice

Mouse work in this chapter was conducted in collaboration with Dr. Chengcheng Jin in the Jacks Lab.

2.1 Abstract

The lung microbiome has a demonstrated impact on the development of asthma and other lung diseases, and the use of culture-independent techniques like sequencing have enabled highly sensitive detection of lung microbes. Human studies are valuable for studying the lung microbiome in the context of disease, but samples are challenging to acquire and potentially contaminated by oral microbiota. Mouse models bypass some of these hurdles by allowing for direct access to the lung through dissection, and the controlled environment, ability to perform perturbation experiments, and different disease models in mice provide an added benefit. However, the low microbial biomass in healthy pulmonary systems and the small size of the mouse lung make accurate sampling and detection of microbes challenging. Here we report a procedure that minimizes PCR cycles and background contamination and provides consistent results across mouse samples. Importantly, we define a negative control panel diagnostic for common problems with small animal lung microbiome studies. We systematically addressed each protocol step from DNA extraction to PCR amplification of the 16S rRNA gene and optimized to maximize sensitivity and minimize sources of contamination for mouse lung microbiome samples. We applied this procedure

to examine the lungs of immunodeficient mice and found that *Rag2*$^{-/-}$ mice have decreased microbial diversity compared to wild-type mice. This result signifies that the immune system may play an active role in modulating the lung microbiome.

2.2 Introduction

The lung harbors a collection of bacteria, in both health and disease. Previously, culture-based methods had failed to detect bacteria in the healthy lung (which in this context we use to mean the lower respiratory tract, consisting of both the trachea and lung). However, the application of DNA sequencing lowered the threshold of detection and enabled identification of bacteria in the lungs. Even though the high amount of mammalian DNA in lung samples can make signal detection challenging, microbial DNA can be identified through the use of 16S rRNA gene sequencing. Studies based on 16S sequencing report that mammalian lungs contain microbial DNA from genera including *Staphylococcus, Streptococcus, Lactobacillus, Rastonia, Enterobacteria, Fusobacteria, Pasteurella,* and *Pseudomonas* [35], [36], [48]–[51].

The lung microbiome has been shown to be associated with the progression of respiratory diseases such as asthma and chronic obstructive pulmonary disease [34], [35], [37], [52]–[54]. Gollwitzer *et al.* showed that microbial signals in the lung of neonatal mice improve immune tolerance to allergens, establishing a clear role of microbes in the lung immune response. Ongoing efforts to understand the relationship between microbes in the lung and effect on disease will rely on animal models, since there are ethical and practical limitations in human studies. Ethics issues associated with research may include invasiveness during sampling, minimizing risk, and proper informed consent. To sample the human lung, clinicians may: 1) collect induced sputum, 2) pass a bronchoscope passed through the mouth and trachea to collect a lavage or lung brushing, or 3) collect lung tissue via sterile surgical explant. Except in the case of invasive surgery, all of these methods require the sample to pass through the upper respiratory tract, potentially collecting contaminating microbes from the oral cavity. In contrast, mouse models provide a system in which lung tissue can be more easily and ethically excised or separated from the upper respiratory tract, minimizing contamination risk. Mouse models enable in-depth examination across disease types and genetic backgrounds, and the controlled housing environment minimizes heterogeneity across samples. Furthermore, mouse models enable the introduction of a perturbation to study specific changes in a controlled environment.

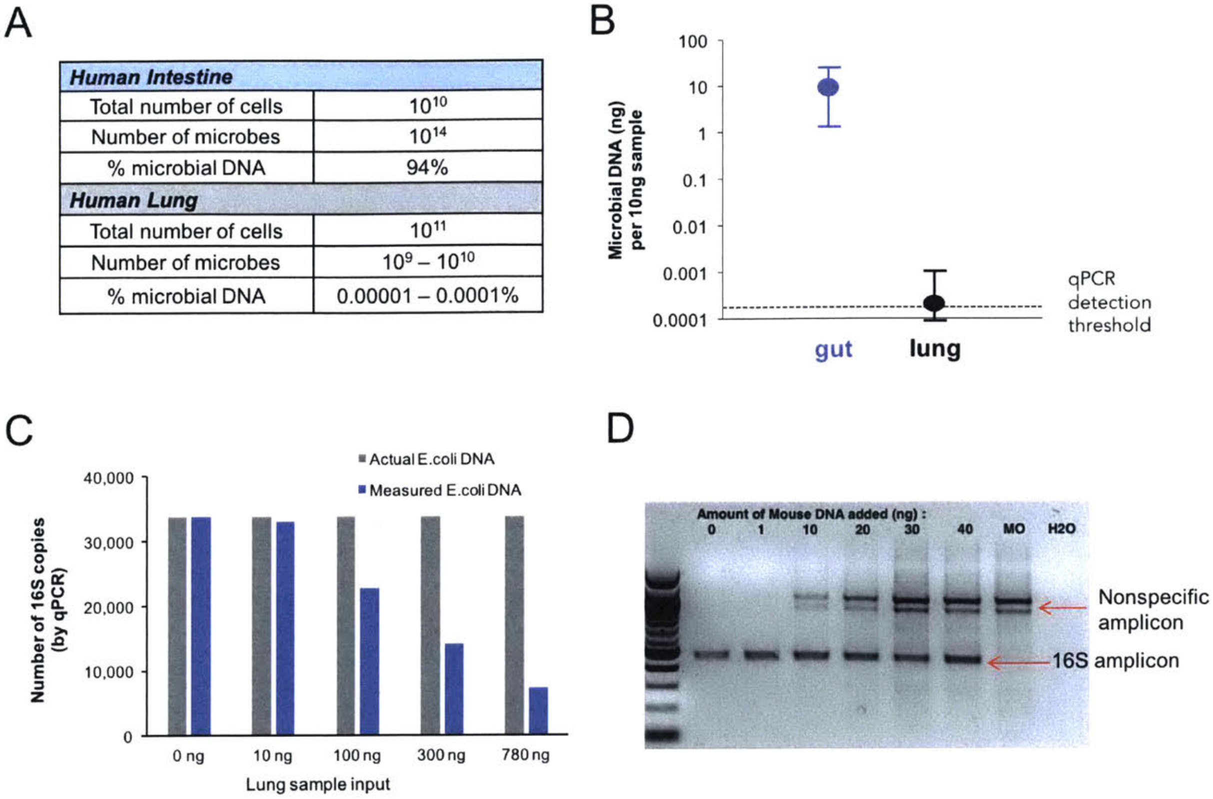
The field of lung microbiome research has substantial technical challenges and several studies report conflicting results. It is not clear whether the lung microbiota is primarily derived from the oral microbiota, as human samples may be contaminated as they pass through the oral cavity upon collection. Furthermore, the use of different reagents and protocols confound result comparison. To date, eight studies have published work on 16S-based sequencing of the mouse lung microbiota [36], [49], [51], [55]–[59]. The work in the lung microbiome field is challenging due to the low microbial biomass in the lung, and many of these groups have utilized large numbers of PCR amplification cycles (>40 cycles). This can introduce bias by distorting the true DNA sequences as well as amplify extremely low levels of contaminating DNA that may be present in samples or reagents. In contrast to a stool sample that contains 99% microbial DNA by mass (out of the combined microbial plus mammalian DNA in one sample), a lung tissue sample has less than 0.0001% microbial DNA (see Figure 2-1). This poses multiple challenges: PCR inhibition from host DNA, off-target amplification, and background noise. We developed a procedure to minimize PCR amplification cycles and background contamination, as well as provide consistent results across mouse samples. We sampled and sequenced over 200 mice, and we applied this method to investigate the microbiota of immunodeficient mice.

We studied two models of immunodeficient mice (*Rag2*$^{-/-}$ and *Tlr2*$^{-/-}$) in order to explore the interaction between the immune system and the lung microbiota. *Rag2*$^{-/-}$ mice are deficient in the recombinase-activating gene 2, which is critical for the development of mature B and T lymphocytes. Thus, Rag2-deficient mice do not have functioning adaptive immunity. Second, we utilized mice deficient in the toll-like receptor 2 gene (TLR2), which is plays a role in detecting microbes and subsequently activating the innate immune system. TLR2 is a surface receptor that recognizes pathogen-associated molecular patterns and helps the immune system determine which bacteria to mount an immune response against [60], [61]. TLR2 recognizes lipoproteins generally from Gram-positive bacteria [62], but also from Gram-negative bacteria [63]. Studies of the gut microbiota in immunodeficient mice have shown some cases of altered communities or susceptibility to infection [64], [65], but little is known about the general impact of innate and adaptive immunity on the lung microbiome.

2.3 Results

2.3.1 Low microbial biomass in lung poses a hurdle to microbiome analysis

In contrast with the gut microbiome, samples from the lung microbiome are at an inherent disadvantage for facile detection of microbial DNA. Host cells outnumber microbial cells by a factor of 10-100 in the lung, while in the gut microbial cells far outnumber host cells (Figure 2-1A). Furthermore, each mammalian cell has 1000-fold more DNA by mass than a microbial cell. Figure 2-1B shows that the estimated amount of microbial DNA in a 10 ng total DNA sample lies just at the qPCR detection threshold (the point at which signal is also detected from no template controls). In order to analyze the microbial DNA, the 16S ribosomal RNA gene can be used as a target for selective bacterial amplification, as this gene is conserved across bacteria but does not appear in mammalian calls. When using primers for the 16S region, however, host mammalian DNA can inhibit the PCR amplification reaction (Figure 2-1C). High concentrations of non-specific DNA in a PCR reaction can prevent primers from binding to their targets. With increasing amounts of lung DNA, we found that a consistent spike-in of *E. coli* has reduced read-out through qPCR. We used this titration experiment to determine our maximum threshold for lung sample input to a PCR reaction as 100 ng per 20uL of reaction. Furthermore, non-specific host amplicons appear when using the 16S primers (Figure 2-1D). Therefore, we defined a set of parameters to test and optimize in order to accurately amplify and sequence microbial DNA in mouse lungs. Table 1 summarizes the steps we optimized in this 16S sample preparation procedure.



Human Intestine	
Total number of cells	10^{10}
Number of microbes	10^{14}
% microbial DNA	94%
Human Lung	
Total number of cells	10^{11}
Number of microbes	$10^{9} - 10^{10}$
% microbial DNA	0.00001 – 0.0001%

Figure 2-1: Challenges in lung microbiome analysis. A) The lung has a low number of bacterial cells compared to intestine. B) The ratio of microbial DNA to host DNA is very low in the lung as compared to the gut. The qPCR detection threshold is determine by the quantity of DNA that is detected by 16S qPCR at the same cycle number as DNA is detected in the negative controls. C) Host lung DNA inhibits PCR amplification of microbial DNA, with increasing amounts of host DNA further inhibiting the reaction. Five lung DNA PCR input amounts were tested with the same quantity "spike-in" of *E. coli* genomic DNA. D) After 40 cycles of PCR, amplicon products show high amounts of off-target DNA (expected amplicon at 450 bp size, lower band). Mouse DNA is titrated into reactions containing the same amount of *E. coli* DNA. Experiment and gel in (D) performed by Dr. Chengcheng Jin in the Jacks Lab.

Parameters	**Optimal PCR condition**
Optimized input amount	100 ng per 20 uL reaction
2-round PCR amplification	16S primers (round 1), sequencing adapters (round 2)
Primers with minimal off-target amplicons	27F/338R 16S primers
Minimal PCR cycles	20 cycles (round 1), 18 cycles (round 2)

Table 1: Optimized DNA library preparation for lung microbiome 16S analysis in mouse BAL.

2.3.2 Optimized procedure for lung microbiome analysis

Through testing four different 16S primer pairs (Supp. Table 2-1), we found that the 27F/338R primers provided the clearest signal with the least amount of off-target signal or negative control signal (Supp. Figure 2-1). We utilized a two-round PCR amplification strategy (Figure 2-2A), based on protocols from previous work [66], [67]. This two-round PCR amplification has two main advantages over the traditional single-round PCR. First, shorter primers can be used that separately target the 16S region and then add sequencing adapters (rather than in one combined primer with single-round PCR), which reduces the amount of off-target binding. Second, the two-round PCR allows for flexibility to use different types of first-round primers with the same second-round barcoding primers. We optimized enzyme type and PCR cycle number, aiming to have fewer than 40 total cycles.

2.3.3 Bronchoalveolar lavage provides superior results compared to lung tissue

We tested both lung tissue and bronchoalveolar lavage (BAL) fluid. Ideally, lung tissue is the most direct source of measuring the lung microbiome - the entire lung tissue is processed and all present bacterial DNA is included. This is in contrast to BAL, which is a thorough wash of the lung interior, collecting cells and DNA present on the surface, but potentially excluding bacteria strongly adhered to or resident inside of the lung tissue. However, through testing we found that host background from lung tissue was too high: tissue samples did not provide consistent results with less than 40 PCR cycles and many sequencing reads were low-quality or host DNA (Figure 2-2B). The BAL samples provided cleaner results due to the lower host background, and therefore we decided to proceed with BAL samples for all future studies. With our resulting protocol applied to mouse BAL, Figure 2-2C demonstrates a clear signal from lung samples by gel electrophoresis. In comparing our initial testing with our final optimized protocol, we were able to reduce total read count from negative controls by roughly 40% (Figure 2-2D).

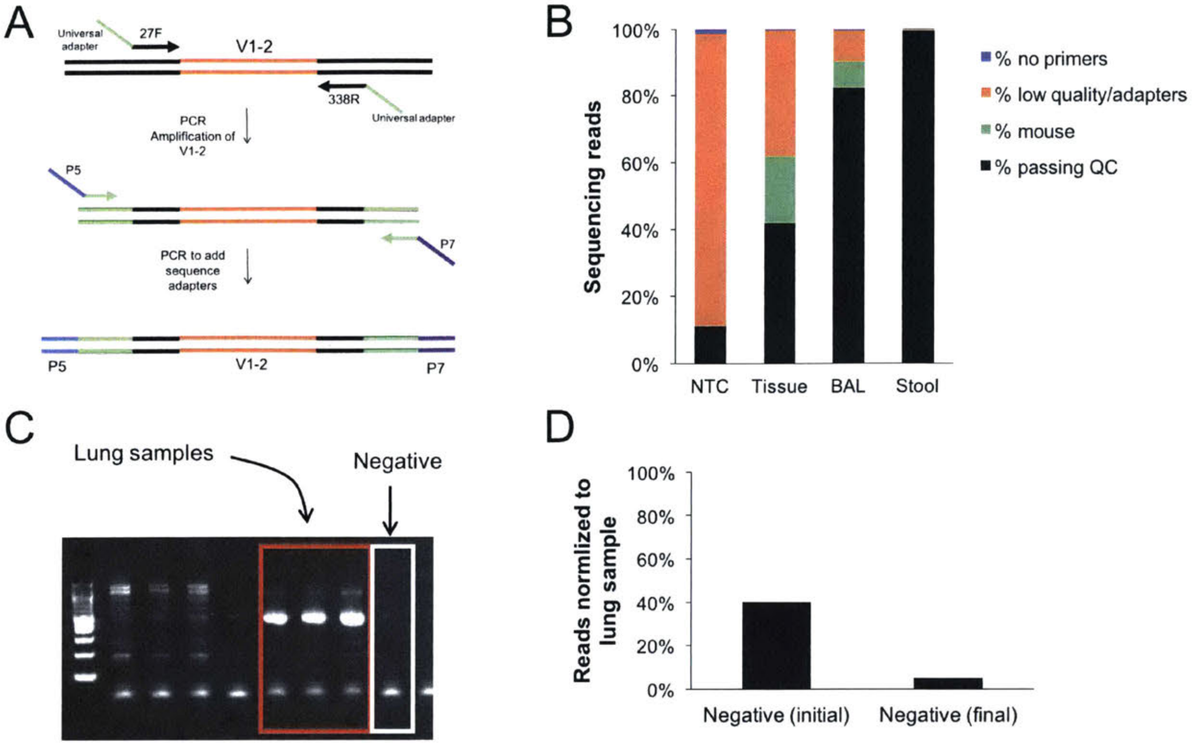


Figure 2-2. Optimized methods to overcome lung microbiome analysis challenges. A) Two subsequent PCR reactions are performed to selectively amplify and tag bacterial DNA. In the first round, 27F and 338R primers for the V1-2 16S variable region are used. The universal adapter at the ends of these primers is then used as the priming site for the second reaction, which adds the Illumina adaptors (P5 and P7) and barcodes (not shown). B) Percent of sequencing reads that do not have the 16S primers, are low quality or adapter primer-dimers, map to the mouse genome, or pass quality control (QC). Samples include PCR no template control (NTC), lung tissue, BAL, and stool (results are the average of 2, 5, 5, and 5 samples, respectively). C) Gel agarose with lung samples (left) and negative control (right) after 38 cycles of amplification with combined two-round PCR. D) Percent of sequencing reads in negative controls normalized by average lung sample reads per experiment for pre-optimized protocol (left) and final protocol (right).

2.3.4 Positive and negative controls enable quality monitoring of lung samples

We applied our 16S PCR protocol to mouse BAL samples. In preparing and sequencing multiple batches of mice, we monitored our negative control samples for signs of contamination. Due to the high total PCR cycle number (38 cycles), negative controls could have potential amplification of contaminant bacterial DNA. Sources could include lab reagents, lab equipment, or any bacteria in the environment (such as the air or skin), and other studies have shown the importance of monitoring contaminants when working with low abundance microbial samples [68]–[70]. To minimize contamination, we set up all 16S PCR reactions in a UV-irradiated cabinet, used UV-sterilized buffers, thoroughly cleaned equipment and gloves with RNAse-away to remove DNA contaminants, and physically separated the set up areas for the first and second PCR reactions. For each experiment we included a set of controls: mock collection control (sterile buffer used in mouse lung collection was placed in tube after completing all mouse collections and went through all procedural steps alongside mouse samples), no-template PCR control (sterile water used for PCR reaction setup), positive PCR control (equal mix of DNA from 20 known bacteria), and triplicate PCR samples for at least one BAL DNA sample. We monitored our mock samples and no-template controls (NTCs) for bacterial sequencing reads. If there was a sign of lab reagent contaminants or systematic protocol contaminants, we would be able to detect this from our control samples.

We completed five different experiments and sequencing runs, and Figure 2-3 shows the normalized number of sequence reads that map to 16S bacterial DNA for each sample (BAL samples are from C57BL/6 mice). Reads were normalized to the median BAL count per sequencing run. We did find that some mock collection blanks had levels of 16S bacterial sequencing reads comparable to BAL samples, while other cases had much fewer. NTCs had on average < 20% of the bacterial 16S reads found in BAL samples of the same sequencing batch.

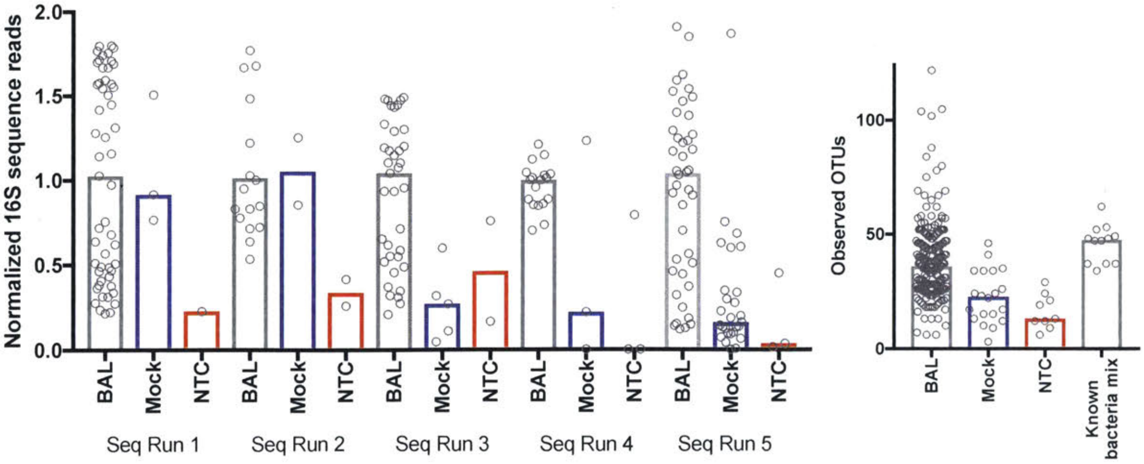


Figure 2-3. Examination of sequencing reads in BAL and controls. Left: normalized 16S sequencing reads for each sample type across sequencing runs. Each experiment had at least one mock collection blank and one PCR no template control (NTC). Right: number of OTUs at >0.1% abundance per sample for BAL, mock, NTC, and known bacteria mix (consisting of 20 types of bacterial DNA mixed in equal proportions). Filtering abundance thresholds were tested with the known bacterial mix to determine 0.1-1% as appropriate threshold (Supp. Figure 2-3).

The positive controls and replicates were used to ensure that sequencing preparation did not have problems and that we were indeed amplifying a reproducible bacterial signal in BAL samples. Replicates were consistent in each batch (Supp. Figure 2-2), indicating that separately prepared BAL samples had the same collection of amplifiable bacterial DNA and providing support that random lab contaminants were not distorting the signal between samples. Positive control samples consistently matched to the known starting bacteria, and no contaminants were found in those samples (Supp. Figure 2-2).

In order to compare the 16S bacterial reads between BAL samples and the negative controls (mocks and NTCs), we examined both the frequency of occurrence and the mean abundance of bacteria in samples. Overall, we found a high prevalence of a few specific bacterial families in the negative control samples (Supp. Table 2-2) – these included *Pseudomonas, Propionibacterium,* and *Comamonadaceae* (classified to the most specific taxonomic level). Each blank sample had on average 3 taxa that were abundantly present (>5% of reads) and most consisted of the top three taxa mentioned. We did not find that negative controls per sequencing run had distinctly different bacteria; rather, they had multiple shared bacteria, indicating there was not batch-specific bias in the negative controls (Supp. Table 2-3).

2.3.5 Data from controls show that some microbes are uniquely found in lung samples

To distinguish which bacteria are potential contaminants and which bacteria are likely derived from the mouse lung, we compared frequency of occurrence of each bacteria (classified to the family level) between BAL and negative controls, as well as mean abundance (Figure 2-4). We find that some bacteria are in a high percent of BAL samples but not in negative controls, indicating a high likelihood of being a true signal from the mouse lung. We applied a simple metric from this data to create a ranked list of bacterial families ordered by likelihood of being a true mouse lung bacteria (Table 2-1). From this ranking, it is clear that a few bacterial families are commonly found in and highly abundant in negative controls; this suggests that these bacteria are likely sources of contamination. However, while we can classify them as common contaminants we cannot exclude them from also being true signal from the mouse BAL.

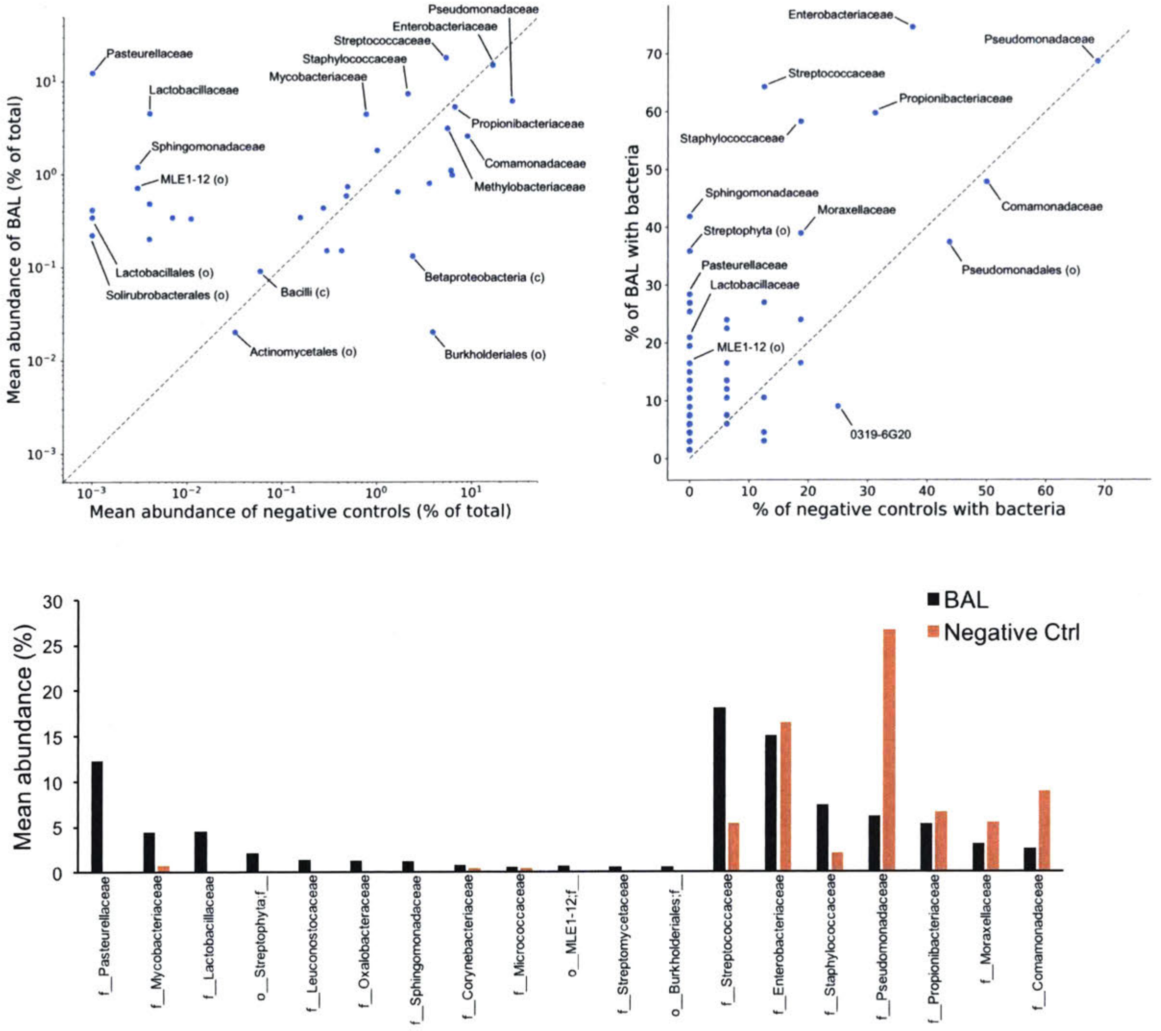
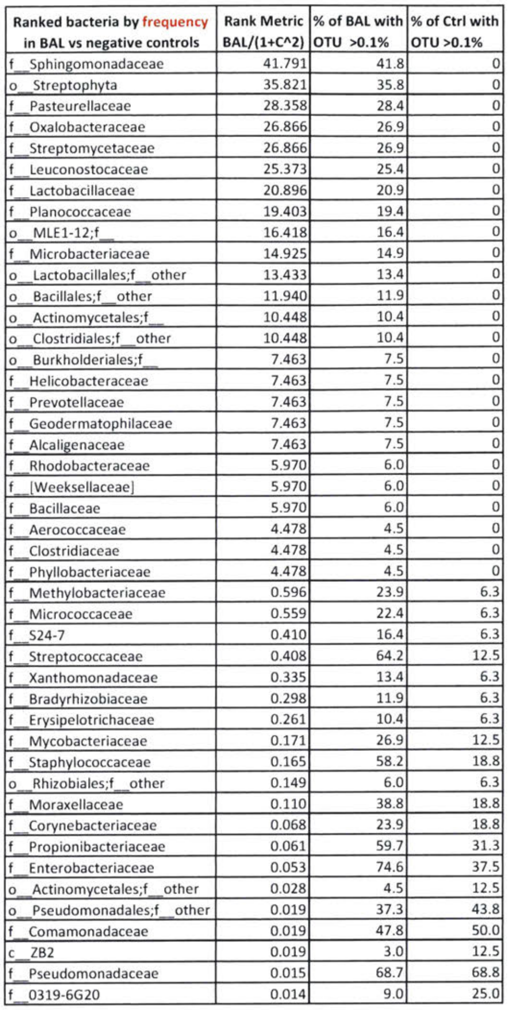
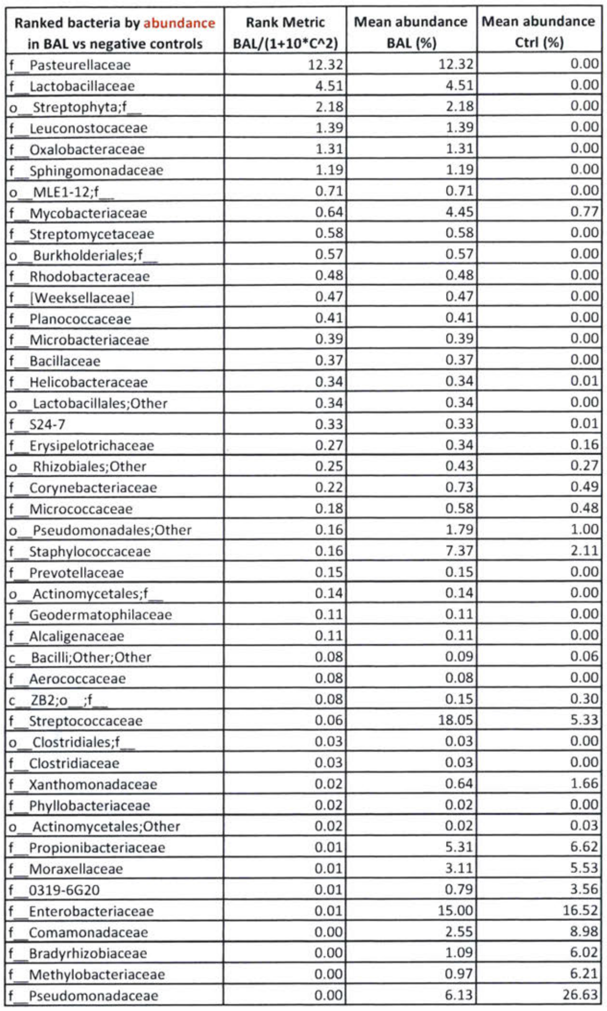


Figure 2-4. Comparison of bacteria found in BAL and negative control samples. A) Mean abundance and B) frequency of bacterial families in BAL versus negative control samples. Bacteria present in the top left represent families that are found in high abundance/frequency in BAL but low abundance/frequency in negative controls. Only a subset of samples are labeled. Due to the log scale in abundance plot, samples with zero abundances in BAL or negative controls do not appear on graph. C) Mean abundance per sample type of top bacterial families found in BAL and in negative controls (f = family, o = order). BAL is black, negative control is orange.



Ranked bacteria by frequency in BAL vs negative controls	Rank Metric BAL/(1+C^2)	% of BAL with OTU >0.1%	% of Ctrl with OTU >0.1%
f__Sphingomonadaceae	41.791	41.8	0
o__Streptophyta	35.821	35.8	0
f__Pasteurellaceae	28.358	28.4	0
f__Oxalobacteraceae	26.866	26.9	0
f__Streptomycetaceae	26.866	26.9	0
f__Leuconostocaceae	25.373	25.4	0
f__Lactobacillaceae	20.896	20.9	0
f__Planococcaceae	19.403	19.4	0
o__MLE1-12;f__	16.418	16.4	0
f__Microbacteriaceae	14.925	14.9	0
o__Lactobacillales;f__other	13.433	13.4	0
o__Bacillales;f__other	11.940	11.9	0
o__Actinomycetales;f__	10.448	10.4	0
o__Clostridiales;f__other	10.448	10.4	0
o__Burkholderiales;f__	7.463	7.5	0
f__Helicobacteraceae	7.463	7.5	0
f__Prevotellaceae	7.463	7.5	0
f__Geodermatophilaceae	7.463	7.5	0
f__Alcaligenaceae	7.463	7.5	0
f__Rhodobacteraceae	5.970	6.0	0
f__[Weeksellaceae]	5.970	6.0	0
f__Bacillaceae	5.970	6.0	0
f__Aerococcaceae	4.478	4.5	0
f__Clostridiaceae	4.478	4.5	0
f__Phyllobacteriaceae	4.478	4.5	0
f__Methylobacteriaceae	0.596	23.9	6.3
f__Micrococcaceae	0.559	22.4	6.3
f__S24-7	0.410	16.4	6.3
f__Streptococcaceae	0.408	64.2	12.5
f__Xanthomonadaceae	0.335	13.4	6.3
f__Bradyrhizobiaceae	0.298	11.9	6.3
f__Erysipelotrichaceae	0.261	10.4	6.3
f__Mycobacteriaceae	0.171	26.9	12.5
f__Staphylococcaceae	0.165	58.2	18.8
o__Rhizobiales;f__other	0.149	6.0	6.3
f__Moraxellaceae	0.110	38.8	18.8
f__Corynebacteriaceae	0.068	23.9	18.8
f__Propionibacteriaceae	0.061	59.7	31.3
f__Enterobacteriaceae	0.053	74.6	37.5
o__Actinomycetales;f__other	0.028	4.5	12.5
o__Pseudomonadales;f__other	0.019	37.3	43.8
f__Comamonadaceae	0.019	47.8	50.0
c__ZB2	0.019	3.0	12.5
f__Pseudomonadaceae	0.015	68.7	68.8
f__0319-6G20	0.014	9.0	25.0



Ranked bacteria by abundance in BAL vs negative controls	Rank Metric BAL/(1+10*C^2)	Mean abundance BAL (%)	Mean abundance Ctrl (%)
f__Pasteurellaceae	12.32	12.32	0.00
f__Lactobacillaceae	4.51	4.51	0.00
o__Streptophyta;f__	2.18	2.18	0.00
f__Leuconostocaceae	1.39	1.39	0.00
f__Oxalobacteraceae	1.31	1.31	0.00
f__Sphingomonadaceae	1.19	1.19	0.00
o__MLE1-12;f__	0.71	0.71	0.00
f__Mycobacteriaceae	0.64	4.45	0.77
f__Streptomycetaceae	0.58	0.58	0.00
o__Burkholderiales;f__	0.57	0.57	0.00
f__Rhodobacteraceae	0.48	0.48	0.00
f__[Weeksellaceae]	0.47	0.47	0.00
f__Planococcaceae	0.41	0.41	0.00
f__Microbacteriaceae	0.39	0.39	0.00
f__Bacillaceae	0.37	0.37	0.00
f__Helicobacteraceae	0.34	0.34	0.01
o__Lactobacillales;Other	0.34	0.34	0.00
f__S24-7	0.33	0.33	0.01
f__Erysipelotrichaceae	0.27	0.34	0.16
o__Rhizobiales;Other	0.25	0.43	0.27
f__Corynebacteriaceae	0.22	0.73	0.49
f__Micrococcaceae	0.18	0.58	0.48
o__Pseudomonadales;Other	0.16	1.79	1.00
f__Staphylococcaceae	0.16	7.37	2.11
f__Prevotellaceae	0.15	0.15	0.00
o__Actinomycetales;f__	0.14	0.14	0.00
f__Geodermatophilaceae	0.11	0.11	0.00
f__Alcaligenaceae	0.11	0.11	0.00
c__Bacilli;Other;Other	0.08	0.09	0.06
f__Aerococcaceae	0.08	0.08	0.00
c__ZB2;o__;f__	0.08	0.15	0.30
f__Streptococcaceae	0.06	18.05	5.33
o__Clostridiales;f__	0.03	0.03	0.00
f__Clostridiaceae	0.03	0.03	0.00
f__Xanthomonadaceae	0.02	0.64	1.66
f__Phyllobacteriaceae	0.02	0.02	0.00
o__Actinomycetales;Other	0.02	0.02	0.03
f__Propionibacteriaceae	0.01	5.31	6.62
f__Moraxellaceae	0.01	3.11	5.53
f__0319-6G20	0.01	0.79	3.56
f__Enterobacteriaceae	0.01	15.00	16.52
f__Comamonadaceae	0.00	2.55	8.98
f__Bradyrhizobiaceae	0.00	1.09	6.02
f__Methylobacteriaceae	0.00	0.97	6.21
f__Pseudomonadaceae	0.00	6.13	26.63

Table 2-1. Ranked list of bacterial families according to likelihood of representing a true signal in mouse BAL. Left: bacteria ranked by frequency in BAL versus negative controls using ranking metric displayed in second column. Right: bacteria ranked by mean abundance in BAL versus negative controls. Bacteria are labeled to the highest resolution taxa (in some cases the sequencing reads are classified only down to order "o" or class "c").

In sequencing BAL samples from C57BL/6 mice, the top bacterial phyla were *Proteobacteria* (44% of total sequence reads), *Firmicutes* (32%), and *Actinobacteria* (15%). If we remove OTUs present in any negative control sample at >1% abundance, then the resulting phylum distribution does not change significantly (Figure 2-5A). After removal of OTUs in negative control samples, the top bacterial families are *Pasteurellaceae*, *Streptococcaceae*, *Staphylococcaceae*, *S24-7*, *Sphingomonadaceae*, *Streptomycetaceae*, and *Lactobacillaceae*, accounting for 55% of total sequencing reads (Figure 2-5B). Some families, such as *Streptococcaceae*, *Staphylococcaceae*, and *S24-7* are ranked low on the likelihood list of true mouse BAL bacterial families but still appear after subtraction. This occurs because some OTUs within the family are present in negative controls while other OTUs are not. According to our ranked list of bacterial families, *Pasteurellaceae* and *Lactobacillaceae* are high-likelihood to be true signal in the mouse BAL.

Previous mouse lung microbiome studies have identified similar phyla (*Firmicutes, Proteobacteria*, and *Actinobacteria*) in the mouse lung [36], [49], [51], [56], [58], [59]. Four of these studies have also found *Lactobacillaceae* and two have found *Pasteurellaceae* as common lung microbes, and many report *Staphylococcus, Corynebacterium, Pseudomonas, Streptococcus, Sphingomonas*, and *Acinetobacter* as additional common lung microbes. Studies have verified the presence of bacteria with bacterial-specific or phylum-specific PCR [35] and with successful culture of bacteria including *Staphylococcus, Streptococcus, Enterococcus*, and *Lactobacillus* [71].

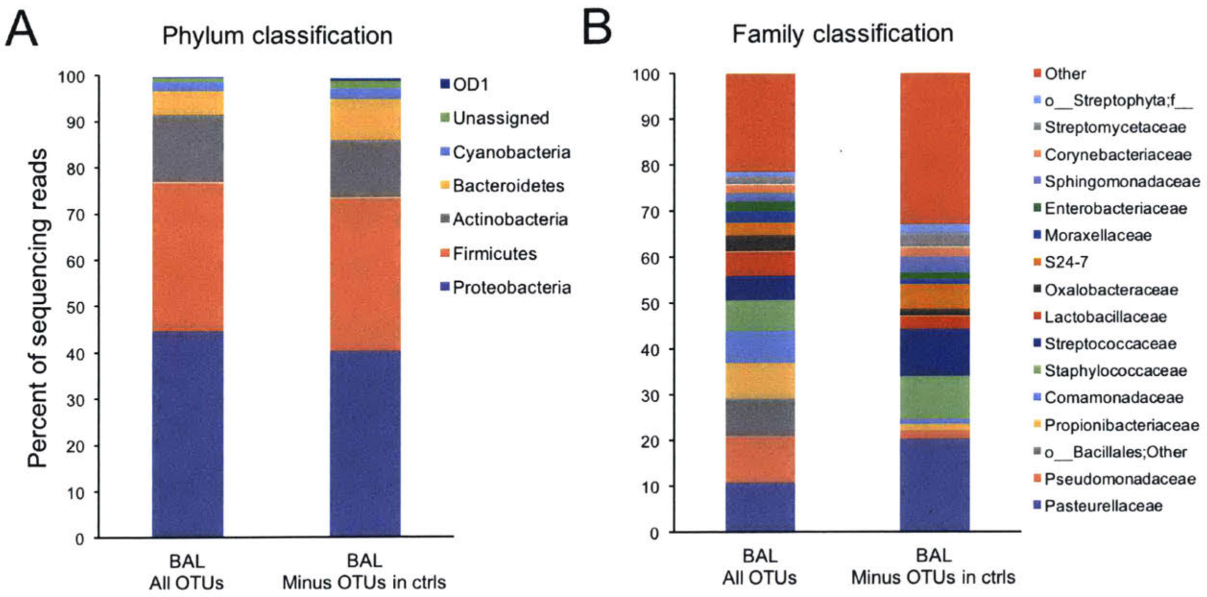


Figure 2-5. Summary-level microbial community in mouse lung. A) Phylum distribution of mouse BAL samples (left, with all OTUs; right, deletion of OTUs found in any negative control samples >1%). B) Family distribution of mouse BAL samples (left, with all OTUs; right, deletion of OTUs found in any negative control samples >1%). Columns represent averaged data from 100 healthy mice.

2.3.6 *Rag2*$^{-/-}$ mice have reduced microbial diversity

To determine how the lung microbiota may change with an altered immune system, we examined mice with a major perturbation of adaptive immunity (*Rag2*$^{-/-}$) and mice with a major perturbation to innate immunity (*Tlr2*$^{-/-}$). Wildtype (WT) mice were sampled along with each cohort. This animal work was performed in collaboration with Dr. Chengcheng Jin in the Jacks Lab. All animals were under the care of the Jacks Lab, and sample collection and DNA extraction was performed both by the author and the Jacks Lab. The identity of the top bacterial families found in the immunodeficient mice compared to wildtype mice is shown in Figure 2-6A. We found that *Rag2*$^{-/-}$ mice were dominated by a single bacterial family, *Pasteurellaceae*, with an average of 95% of total reads. This effect was striking and observed in all animals (from two separate experimental cohorts), and the abundance was significantly different from control animals (Figure 2-6B). No other set of mouse samples demonstrated such a high abundance of a single taxon, particularly one that was not found in any negative controls. *Tlr2*$^{-/-}$ mice had a more diverse community, with *Streptococcaceae, Mycobacteriaceae*, and *Enterobacteriaceae* as the top families. Overall, the wildtype mice had many shared families but high variability between samples. We were surprised to identify *Streptophyta*, a member of the Cyanobacteria phylum. However, this family has been found in multiple studies of the gut microbiome, skin microbiome, and indoor microbiome [72]–[76]. Whole genome analysis of Cyanobacteria from the gut microbiome has suggested some of these bacteria may be a unique lineage of non-photosynthetic organisms that are true residents of the commensal microbiome [77].

To compare the microbial diversity of between sample types, we calculated the Shannon diversity index and found it was significantly reduced in *Rag2*$^{-/-}$ mice compared to controls (Figure 2-6C), while *Tlr2*$^{-/-}$ mice showed no significant change. This reduced diversity does not occur due to reduced microbial load as measured by qPCR, for both WT and *Rag2*$^{-/-}$ have similar amounts of bacteria in the BAL samples. Furthermore, the relative abundance of *Pasteurellaceae* was significantly greater in *Rag2*$^{-/-}$ as compared to controls.

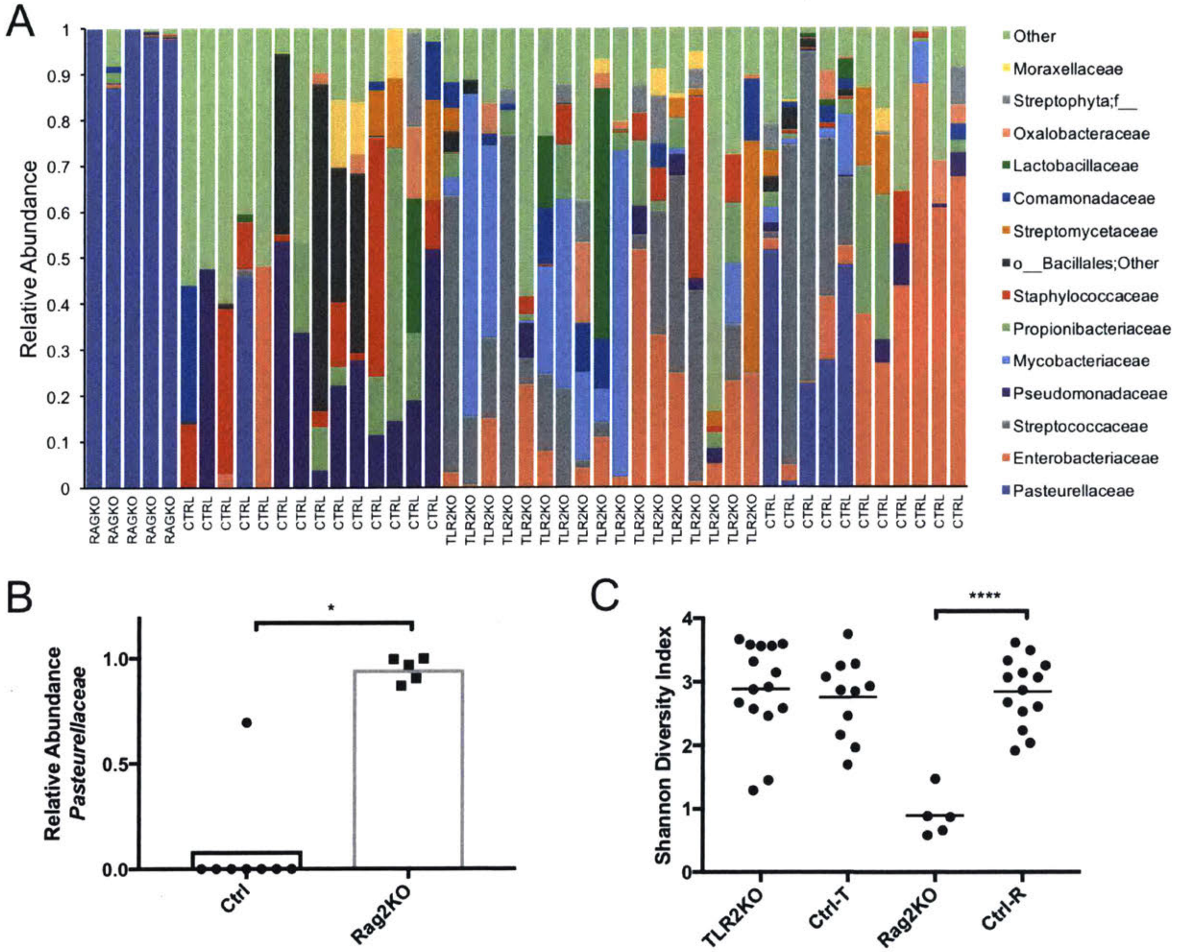


Figure 2-6. Microbial community and diversity in *Rag2*$^{-/-}$, *Tlr2*$^{-/-}$, and wildtype controls. A) Relative abundance of bacterial families. B) *Pasteurellaceae* is significantly increased in *Rag2*$^{-/-}$ mice compared to controls, $p = 0.04$ (family level comparison with Mann-Whitney test, Bonferroni p-value correction with n=25). C) Shannon diversity index is reduced in *Rag2*$^{-/-}$ compared to wildtype controls, $p < 0.0001$ (two-tailed unpaired Student's t-test).

2.4 Discussion

In this study, we developed an optimized protocol for analysis of the low biomass mouse lung microbiome, specifically examining signal in negative controls to identify unique signal from mouse BAL, and we applied these methods to study a cohort of healthy and immunodeficient mice. The lung is constantly exposed to microbes through inhalation and microaspiration of oral and upper respiratory microbes, and the application of sequencing methods to study the lung has identified microbial DNA in the lung and challenged the belief that the lung airways are sterile under healthy conditions.

The lung microbiome presents a challenging sample type for analysis due to the low abundance of bacterial DNA and high abundance of host DNA. Mammalian DNA far outnumbers bacterial DNA mass in the lung by a factor of approximately 10,000. PCR inhibition from host lung DNA, off-target amplification products, and high potential for contamination pose multiple hurdles. Microbial DNA contaminants from lab reagents and the lab environment can easily distort the lung microbiome signal without careful precautions. While a few groups have reported on the mouse lung microbiome with 16S sequencing [36], [49], [51], [55]–[59], not all included adequate controls or reported on the sequences found in negative controls. Over half of the studies have PCR amplification cycles that surpass 40, a high number that signals that the true bacterial DNA may be distorted. And there is no systematic testing or consensus on sample type nor 16S variable region for sequencing.

In our study, we work with mice due to the advantages they have in early stage research, as they provide a model system to study the lung microbiome in a controlled environment and with different perturbations. Since the field is still developing and testing different methods for lung microbiome sampling, the mouse provides an excellent model system to optimize protocols. In order to sample the human lung, sputum samples or collection equipment must pass through upper respiratory tract, introducing the potential for bacterial contamination. Although a recent study has carefully tested for contamination in human lung sample collection [78], the mouse presents a model system that can bypass contamination challenges as well as provide a controlled environment to study various conditions.

In order to progress the field of mouse lung microbiome research, it is necessary to test sample type and 16S PCR conditions to optimize a protocol for lung microbiome analysis. We demonstrated that bronchoalveolar lavage (BAL) fluid was superior to lung tissue by minimizing off target PCR products and producing a more clear bacterial signal through 16S sequencing. This does not preclude lung tissue samples from future microbial 16S sequencing, but we believe that further protocol optimization or host DNA depletion methods are required to generate a clean, reliable signal [79]–[81]. It is important to note that while the lavage has less host DNA, this method of collection may provide a different signal compared to direct tissue sampling. We systematically tested conditions for 16S PCR amplification to enable sensitive detection of bacterial DNA while minimizing contaminants. We found that the V1-V2 16S variable region provided the clearest signal in mouse BAL, minimizing off target amplicons and allowing for the fewest PCR cycles.

We tested a panel of negative and positive controls along with our mouse BAL samples to determine any signs of contaminants. Our 16S sequencing preparation protocol produced consistent results with replicate sample preparation, indicating that bacteria can be reproducibly amplified from BAL samples. Although we did find bacterial DNA signal in our negative controls, we were able to rank bacteria indicating their likelihood to be a true mouse BAL signal. We demonstrated that a few bacterial families are commonly found in negative control samples (*Pseudomonadaceae, Propionibacteriaceae,* and *Comamonadaceae)*, while other are present only in BAL samples (*Pasteurellaceae* and *Lactobacillaceae*, among others). Similar to previous studies of human and mouse lung microbiome, we found that the dominate phyla were *Firmicutes, Proteobacteria, Actinobacteria,* and *Bacteroidetes,* consisting of 90% of the total sequencing reads.

Previous groups have conflicting results on the mouse lung microbiome and negative controls. Some groups have identified a core microbiome [49], [56], [58], others have found no common microbiota across samples [55], and still others have found that only a fraction of samples have a unique signal that is truly the lung microbiome as opposed to contaminant or upper respiratory tract bacterial DNA [57]. Some of these studies included negative controls, but many do not report on the sequencing results. Researchers either subtracted all OTUs present in controls or did not sequence negative controls. We felt it was most appropriate to present the full data and acknowledge that the lung microbiome is an inherently noisy sample type with the potential for contamination. It is difficult to have absolute certainty over which samples are truly present in the lung versus contaminant samples from reagents or the environment. Many of the microbes detected in the lung are also common environmental microbes that are found in the air and on the skin. Therefore, it is necessary to interpret results cautiously and recognize that "contaminant" bacterial DNA may be found in controls but could also be a true signal in the lung. Furthermore, clear variation due to animal vendor, housing facility, caging, gender, and age confounds results and makes comparison across studies challenging [50], [55], [56]. A major hurdle in validating sequencing results is the difficulty in culturing bacteria from the lung. Few groups have been successful with lung microbe culture from SPF mice, and we have also had multiple failed attempts. Further progress in validating lung microbes is necessary, whether through culture, fluorescence *in situ* hybridization staining of bacteria in lung tissue, or bacteria-specific PCR.

We tested the BAL of 25 healthy mice and 22 immunodeficient mice (*Rag2*$^{-/-}$ and *Tlr2*$^{-/-}$). The lung microbiome of immunodeficient mice has not been previously examined, although there is a clear relationship between the immune system and the lung microbiome. Murine studies have shown that the early lung microbiota promotes tolerance to allergens via PD-1/PD-L1 signaling in regulatory T cells [35] and that specific bacterial strains modulate lung asthma susceptibility [71]. We found that the lung microbiome of *Rag2*$^{-/-}$ mice was overwhelmingly dominated by *Pasteurellaceae*. This finding was unexpected, given that no other mouse types showed dominance of a single taxon and suggests that the perturbation in the adaptive immunity in *Rag2*$^{-/-}$ mice played a role in this *Pasteurellaceae* dominance. *Pasteurellaceae* is a large family of gram-negative bacteria that are mostly known as commensals and are commonly found on mucosal surfaces in the upper respiratory tract. Multiple bacteria in this family are opportunistic pathogens (such as the genera *Haemophilus* and *Pasteurella*), normally coexisting with the host but in certain cases acting as invaders [82]. This coincides well with our finding of *Pasteurellaceae* dominance in the adaptive immune deficient mice, indicating that this bacteria may be able to overcome the host defense mechanisms and outgrow other bacteria under favorable conditions. *Pasteurellaceae* have traits that confer general resistance to cell defense mechanisms, such as an antiphagocytic capsule and exotoxin secretion to inhibit leukocyte antimicrobial activity [83], [84]. This bacterial family was also found in the lungs of wild-caught mice, but did not appear in conventionally reared mice from specific-pathogen free facilities [36]. This may be because of strict controls in the SPF facility, which generally try to screen for and limit detectable amounts of *Pasteurella pneumotropica,* given that it is an opportunistic pathogen. *Tlr2*$^{-/-}$ mice did not show a distinct bacterial community difference compared to controls.

The source of the lung microbiome is still under investigation, as some studies have contributed contrasting evidence [57], [85]. However, it is very likely that the upper respiratory tract has a role in seeding the lung microbiota and there is a balance between intake of bacteria and elimination through cilia clearance, coughing, and local host defenses [30], [78]. It is important to note that while sequencing can identify what bacterial DNA is present in the lung, it does not indicate if the bacteria is alive or dead. It is likely that some of the lung bacterial DNA is extracellular or from dead microbes, but nevertheless there is clear evidence that direct microbial exposure in the lung plays a role in shaping the immune system [35]. Furthermore, our finding of *Pasteurellaceae* abundance in the *Rag2*$^{-/-}$ mice is evidence for some fraction of the lung microbiome to be a growing and replicating community.

Our study had a number of limitations. Given the existing evidence of the effect from animal vendors and housing facilities, it would be ideal to test our *Rag2*$^{-/-}$ finding across multiple mouse types from different vendors or in different facilities. Furthermore, we did not collect nasal rinses or oral swabs to compare our BAL sequencing results with the upper respiratory tract, but we did follow a collection protocol that minimized any contaminants from the upper airway. Additionally, we found the mouse lung microbiome generally shared common taxa but relative abundances were quite variable between samples. This variability can make it difficult to generalize results and compare specific bacterial taxa with other studies. However, the methods presented here can be applied to other samples and ranked list of bacteria can be compared.

In summary, we have developed an optimized protocol for mouse lung microbiome analysis and determined that BAL samples are most ideal for bacterial DNA sequencing. Here we present all sequencing data from negative controls and propose a ranked list of bacteria likely to be found in the mouse lung, given their abundance in either BAL or negative control samples. In future lung microbiome studies, we believe it is critical to have a careful set of positive and negative controls, verifying both the technical reproducibility of the procedure as well as tracking sources of contamination. We call for future researchers to include a negative collection control, PCR control, positive sample control with a known bacterial mix, and multiple technical replicates. Making this a new standard would enable comparison of negative control data and procedural reproducibility across studies. In applying our methods, we found bacteria that were uniquely present in mouse lung samples, including *Pasteurellaceae*, a bacterial family that was in high relative abundance in *Rag2*$^{-/-}$ mice. Future work may investigate what specific role adaptive immunity has in lung microbial homeostasis.

2.5 Methods

2.5.1 Animal work

Mice were housed at the Koch Institute in a specific-pathogen-free environment under care of the Jacks Lab. Food and water was sterilized, and each cage had an individual HEPA filter. The mice were supplied with water and food ad libidum, maintained with a 12-hour light cycle, kept at temperatures between 68–72°F. All mice were on the C57BL/6 background, including *Tlr2*$^{-/-}$ and *Rag2*$^{-/-}$ mice, and mice were bred from in-house heterozygous breedings. All studies were performed under an Institutional Animal Care and Use Committee and Massachusetts Institute of Technology Committee on Animal Care-approved animal protocol. Mice were assessed for morbidity according to MIT Division of Comparative Medicine guidelines and were always humanely sacrificed prior to natural expiration.

2.5.2 Sample collection

The type of samples collected included whole lung tissue, lung lavage, or control samples (blank lavage liquid). The Jacks Lab provided all mice for experiments, and the mouse sample collection was performed in collaboration with the Jacks Lab. Mice were euthanized with isoflurane overdose following proper procedures. After confirmation of death, the exterior of the mouse was sprayed with 70% ethanol and wiped down. Animals were surgically dissected for collection of lung tissue or bronchoalveolar lavage (BAL). First, one set of sterile instruments was used to remove the skin to expose the upper thoracic cavity. A second set of sterile instruments was used to cut open the thoracic cavity and expose the lung and trachea. In the case of lung tissue collection, the whole lung tissue was excised, cutting at the base of the trachea immediately above the bifurcation point at the lung. The lung tissue was placed in sterile tubes containing RNAlater and placed on ice during the rest of the sample collection, after which all tubes were placed at -80C.

In the case of BAL collection, the mouse was pinned to a Styrofoam board to align the head and body in a straight line. A catheter was inserted in the bottom half of the trachea, pointing down towards the lung.

The needle was removed and replaced with a 1mL syringe filled with 1mL of sterile PBS. Upon verification of a tight seal at the insertion point, PBS was slowly pushed into the lungs and then pulled back out into the syringe. This fluid was deposited into an eppendorf tube, and the lavage was repeated once more with another 1mL of PBS. The resulting 1.5 – 2mL of sample was stored on ice during the remainder of the sample collection. After each mouse collection, the collection area and surgical tools were decontaminated with ethanol. At the end of each mouse cohort collection, a blank control sample was collected. PBS solution was pulled into a 1mL syringe and deposited into an empty eppendorf tube, using the same tools and reagents. For storage of BAL samples, immediately after collection all samples were spun down at 10,000 g for 20 minutes at 4C. The supernatant was removed, leaving a small amount of liquid and the cell pellet at the bottom of the tube. Samples were stored at -80C.

2.5.3 DNA extraction

Multiple DNA extraction methods were tested in collaboration with the Jacks Lab. The most successful method for high DNA yield and minimal contaminants was a modified phenol-chloroform extraction combined with bead-beating and chemical lysis based on Turnbaugh et al., adapted from the Goodman Lab and the Jacks Lab [86]. First, cell lysis was performed with bead-beating and addition of detergent. The BAL pellet was resuspended in 400µL sterile PBS and transferred to a 2mL screwtop tube. The following were added to the sample: 250µL of 0.1-mm diameter zirconia/silica beads (BioSpec Products), 300µL of extraction buffer (200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA), 200µL 20% SDS, and 500µL of phenol:chloroform:isoamyl alcohol (25:24:1, pH 7.9). Cells were then mechanically lysed with a bead beater (BioSpec) for 2 minutes, followed by incubation on ice for 2 minutes, and one additional round of bead beating for 2 minutes. Samples were centrifuged at 6,000 g at 4C for 3 minutes. Next, DNA was isolated with phenol-chloroform extraction. The top aqueous phase (~800µL) was transferred from each sample tube to a new sterile eppendorf 2mL tube. An equal amount (~800µL) of phenol:chloroform:isoamyl alcohol was added and mixed by inversion 15 times. Tubes were centrifuged for 5 minutes at 20,000 g at room temperature. The top aqueous phase was transferred to a new tube, and 3 M NaOAc (pH 5.5) was added to achieve a 10% increase in total volume. The solution was mixed thoroughly by vortexing, and subsequently 100% isopropanol (cold, stored at -20C) was added at equal volume. The sample was stored at -20C for at least 2 hours to chill and then centrifuged at 20,000 g at 4C for 20 minutes. The supernatant was removed, keeping the DNA pellet. The pellet was washed by adding 950µL of 100% ethanol and centrifuged for 10 minutes at 4C. Finally, the supernatant was carefully removed and the tube cap was kept open to air dry the DNA pellet. After drying, 50µL buffer (10 mM Tris-HCl, 1mM EDTA, pH7) was added to the pellet and incubated at 50C for 30 minutes or until dissolved, vortexing every 10 minutes. DNA samples were quantified with Nanodrop and normalized to 100 ng/µL with sterile TE buffer (10 mM Tris-HCl, 1mM EDTA, pH7). Samples were stored at -20C for use within a few weeks or moved to -80C for long term storage.

To extract DNA from lung tissue samples, the samples were processed with the PowerSoil DNA Isolation kit (Mo Bio) following the manufacturer's instructions. Other extraction protocols tested with lung samples (tissue and BAL) included the QiaAmp blood and tissue kit and the Qiagen Microbiome kit. Through testing all kits alongside the phenol-chloroform method, we found that the phenol-chloroform gave more consistent results, higher DNA yield from BAL, and the cleanest amplicons (as determined by agarose gel).

2.5.4 qPCR to quantify microbial DNA

In order to measure the amount of microbial DNA, qPCR was performed with the primers 27F (AGAGTTTGATCMTGGCTCAG) and 338R (TGCTGCCTCCCGTAGGAGT) that target the V1-2 region of the 16S rRNA gene. 100ng of DNA was added to 10µL Kapa SYBR Fast 2X master mix (Kapa Biosystems), 1µL 10uM forward primer, 1µL 10uM reverse primer, 1µL Eva dye, 0.2µL Rox reference dye, and PCR-clean water to bring the total reaction volume up to 20µL. The reaction was then run at 95C for 3 minutes; 40 cycles of 95C for 10 seconds, 57C for 20 seconds, and 72C for 60 seconds; and 72C for 1 minute. *E. coli* genomic DNA was used as a standard, with samples at 1, 0.1, 0.01, and 0.001 ng per reaction used to create a standard curve.

2.5.5 16S amplification and sequencing

The 16S rRNA gene was amplified from extracted DNA and used for sequencing. We tested multiple primer pairs targeting different 16S variable regions: V1-2 region with 27F/338R, V4 region with

515F/806R, V3-4 region with 340F/772R, and V6 region with 967F/1061R (Supp. Table 2-1). We utilized two rounds of PCR to prepare sequencing libraries: the first PCR amplified the 16S region and added adapters at each end, and the second PCR added full sequencing adapters and barcodes. As described in the results, the two-round PCR reaction was performed to 1) allow for shorter primers that minimized off-target binding and 2) enable testing of multiple 16S primer pairs that were all compatible with the same barcoding primers in the second PCR reaction. These primers and protocol were adapted from previous work [66], [67]. See diagram in Results section for schematic of two-round PCR.

For the first PCR reaction, 1µL of 100ng/µL of DNA was added to the following: 10µL Kapa SYBR Fast 2X master mix (no dye added, Kapa Biosystems), 1µL 10uM Step1 forward primer, 1µL 10uM Step1 reverse primer, and 7µL PCR-clean water. The 20µL reaction mix was divided into 2 separate tubes with 10µL each for PCR cycling, using the following conditions: 95C for 3 minutes; 20 cycles of 95C for 10 seconds, 57C for 20 seconds, 72C for 60 seconds; and 72C for 1 minute. To purify the PCR product, Agencourt AMPure XP beads (Beckman Coulter) were used following the online protocol. Briefly, the 10µL individual reactions were pooled to yield 20µL total. Equal volume of beads was added to the PCR products, incubated for 5 minutes, and placed on a magnet for 2 minutes to allow for magnetic bead separation. The clear solution was discarded, keeping the beads intact, and each sample was washed twice with 150µL 80% ethanol. Samples were dried for 5 minutes and DNA was eluted with 20µL 10mM Tris-HCl. After a 2 minute incubation, the sample was again placed on the magnet and the clean DNA solution was transferred to a new tube.

To perform the second PCR reaction, 8µL of DNA product from the first reaction was added to the following: 10µL Kapa HiFi HotStart 2x ReadyMix (Kapa Biosystems), 1µL 10uM Step2 forward primer, and 1µL 10uM Step2 reverse primer. The PCR cycling conditions were 95C for 3 minutes; 18 cycles of 98C for 20 seconds, 80C for 15 seconds, and 72C for 60 seconds; and 72C for 3 minutes. A second bead purification was performed to clean the PCR products. The final sample was quantified using the Qubit DNA assay (Thermo Fisher Scientific), and all samples were normalized and pooled at equal mass for the final sequencing pool. Samples were visualized on a 1% agarose gel for appropriate fragment size (approximately 450bp). Samples were sequenced on the Illumina MiSeq with 2 x 250 bp paired-end reads.

2.5.6 Sequence analysis

Sequencing reads were demultiplexed with *bcl2fastq* (Illumina software). FASTQ paired end reads were joined using *multiple_join_paired_ends.py* from QIIME version 1.9.1 with a minimum overlap of 20 bp and maximum percent difference of 40% [87]. Reads that did not contain both forward and reverse primers were removed and primers were trimmed from the start/end of the remaining reads using cutadapt [88]. Trimmomatic was used to do further quality filtering and remove reads that had a Q-score less than 15 in a 4 bp sliding window and remove reads that had sequencing adapters or primers in the middle of the read [89]. Then *multiple_split_libraries_fastq.py* was used in QIIME to do final filtering and apply sample labels to each sequence. This step filtered out sequences less than 200 bp in length or with more than 6 ambiguous bases (along with other default settings). As multiple sequencing runs were performed, these filtering steps were conducted on reads from a single run and subsequently combined together into a single FASTA file.

OTUs were determined from combined sequences using QIIME *pick_open_reference_otus.py*, with Greengenes 13.8 serving as the reference database for both OTUs and taxonomy [90]. Sequences were clustered into 97% identity OTUs and assigned taxonomy using UCLUST [91]. OTUs were assigned taxonomy to the deepest taxonomic level with a minimum 90% similarity from the Greengenes database. OTUs that did not appear in at least 3 samples and have at least 50 sequence counts were removed. On average, BAL samples had 50,000 reads, and a minimum threshold of 500 reads after filtering was used to retain samples.

To calculate diversity metrics, the QIIME pipeline was used. For alpha diversity, we calculated the Shannon index of the 97% identity OTUs using *alph_diversity.py*, with samples rarefied to 1000 reads. Beta diversity measurements were calculated using *beta_diversity.py* to determine the Bray-Curtis, weighted UniFrac, and unweighted UniFrac metrics. PCoA plots were made using *make_2d_plots.py*.

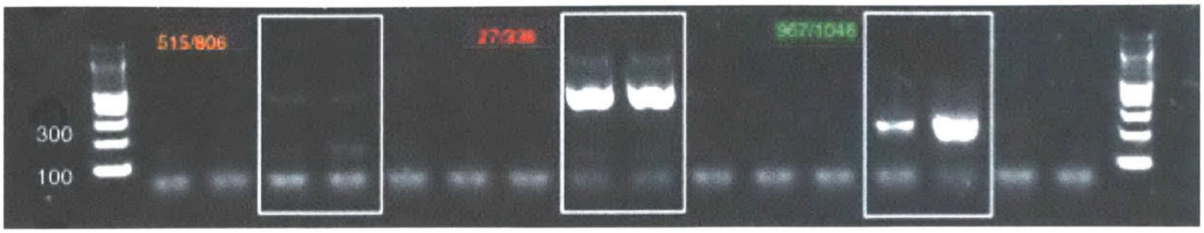
2.5.7 Statistical analysis

We utilized the Student t-test in GraphPad Prism 7 to compare Shannon diversity indices (with significance for p<0.05). To test if bacteria were significantly altered in abundance between cohorts, the non-parametric Mann-Whitney U test was used to compare the ranks of bacterial abundance per sample, with a Bonferroni correction for multiple hypothesis testing (using QIIME *group_significance.py*). Testing was done at the OTU level as well as each taxonomic level with relative abundance counts, with p values corrected according to the number of categories. To test for differences in community composition between sample groups, we performed Permutational Multivariate Analysis of Variance (PERMANOVA) based on the Bray-Curtis and weighted UniFrac distance matrices using QIIME *compare_categories.py.* Statistical significance between groups was determined using 1000 permutations of sample labels.

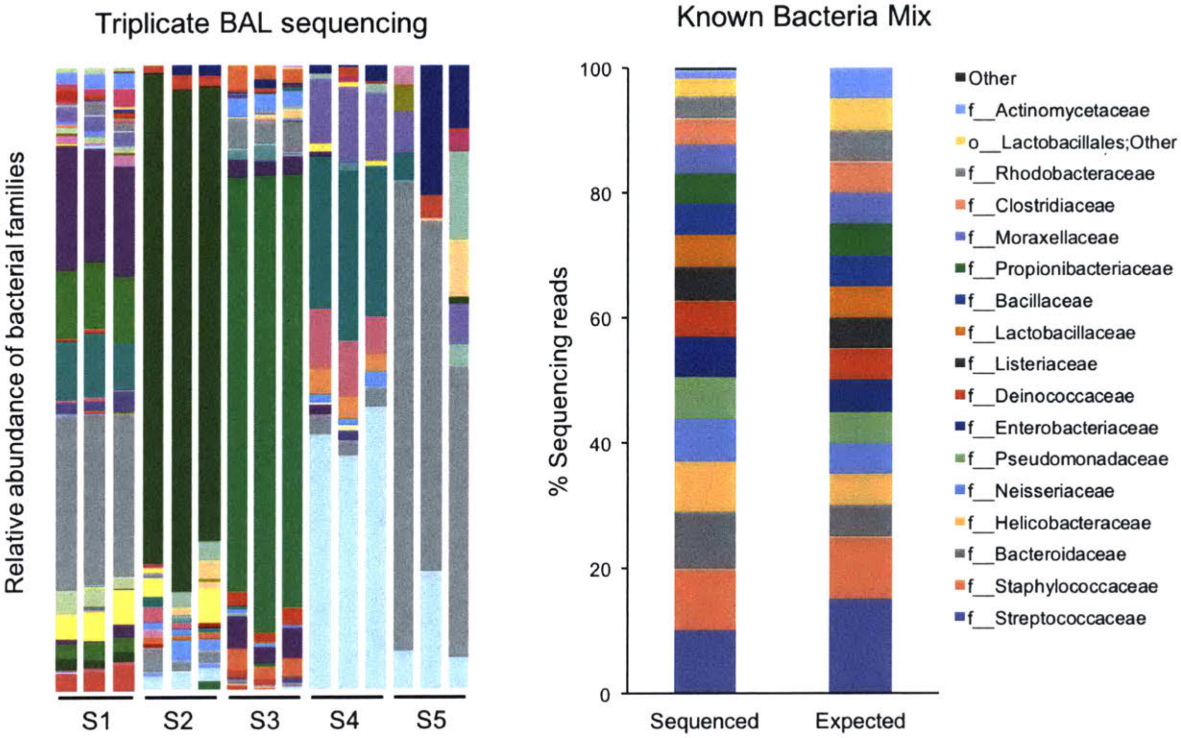
2.6 Supplemental Figures

Primer name	Sequence
Step1 27F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT YRYR AGAGTTTGATCMTGGCTCAG
Step1 338R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TGCTGCCTCCCGTAGGAGT
Step1 967F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT YRYR CAACGCGAAGAACCTTACC
Step1 1061R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CACGRCACGAGCTGACGAC
Step1 340F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT YR TCCTACGGGAGGCAGCAGT
Step1 772R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT GGACTACCAGGGTATCTAATCCTGTT
Step1 515F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT YRYR GTGCCAGCMGCCGCGGTAA
Step1 806R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT GGACTACHVGGGTWTCTAAT
Step2F	AATGATACGGCGACCACCGAGATCT CAC NNNNNNNN ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Step2R	CAAGCAGAAGACGGCATACGAGAT NNNNNNNN GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

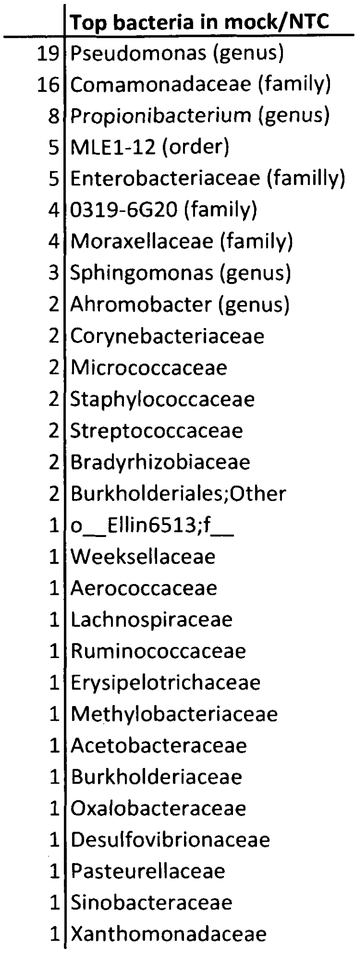
Supp. Table 2-1. 16S rRNA PCR library preparation primers. Step1 primers are used in the first PCR reaction and Step2 primers are used for the second barcoding PCR reaction. Green indicates Illumina sequencing read 1 primer region, blue indicates Illumina sequencing read 2 primer region, orange indicates 16S primer sequence, and underlined indicates unique barcode for sequencing (Ns are replaced with 8bp unique barcode sequence).



Supp. Figure 2-1. Testing of different 16S primers. Gel electrophoresis results from 30 cycles with 515/806, 27/338, or 967/1061 primer pairs (see Supp. Table 2-1). Boxes indicate lung samples with bacteria. Expected product size is about 400bp (515/806), 450bp (27/338), and 200bp (967/1061).

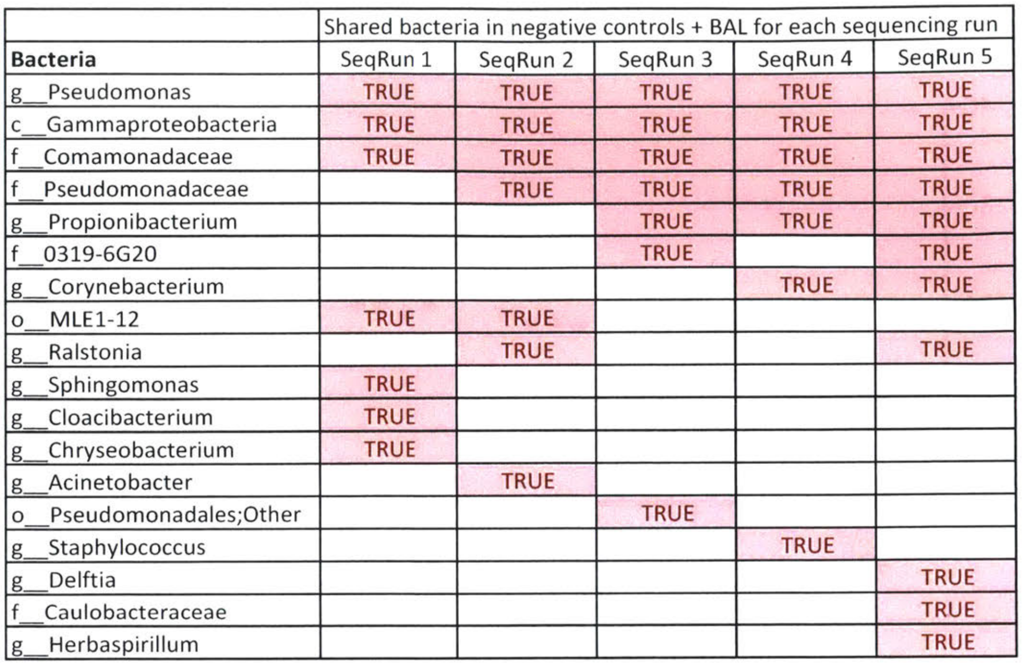


Supp. Figure 2-2. Positive controls samples for BAL sequencing preparation. Left: Triplicate sequencing preparation of mouse BAL samples shows high degree of 16S sequencing reproducibility. Right: known mix of bacterial DNA sequenced as positive control. "Sequenced" bar represents the bacterial taxa identified in the sequenced samples (averaged of all samples) and "Expected" bar represents the exact proportions of bacterial DNA used to create the mixture. No contaminants or unexpected sequences detected in the positive control.



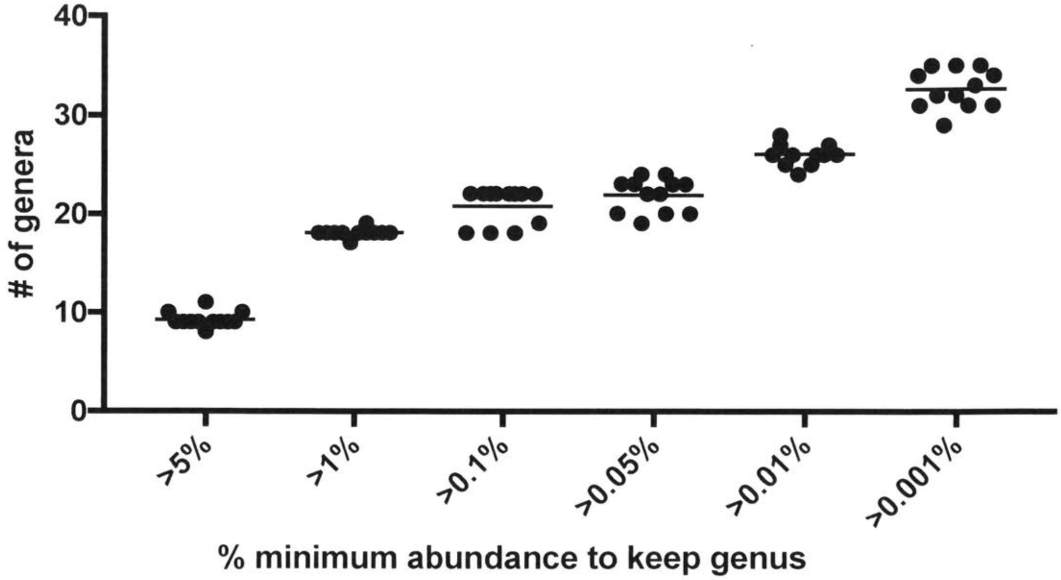
	Top bacteria in mock/NTC
19	Pseudomonas (genus)
16	Comamonadaceae (family)
8	Propionibacterium (genus)
5	MLE1-12 (order)
5	Enterobacteriaceae (familly)
4	0319-6G20 (family)
4	Moraxellaceae (family)
3	Sphingomonas (genus)
2	Ahromobacter (genus)
2	Corynebacteriaceae
2	Micrococcaceae
2	Staphylococcaceae
2	Streptococcaceae
2	Bradyrhizobiaceae
2	Burkholderiales;Other
1	o__Ellin6513;f__
1	Weeksellaceae
1	Aerococcaceae
1	Lachnospiraceae
1	Ruminococcaceae
1	Erysipelotrichaceae
1	Methylobacteriaceae
1	Acetobacteraceae
1	Burkholderiaceae
1	Oxalobacteraceae
1	Desulfovibrionaceae
1	Pasteurellaceae
1	Sinobacteraceae
1	Xanthomonadaceae

Supp. Table 2-2. Taxa found in negative control samples (mock collection or no-template PCR controls). List is ranked by number of times present in mock/NTC samples at >10 sequence counts (total of 24 blank samples).



	Shared bacteria in negative controls + BAL for each sequencing run				
Bacteria	SeqRun 1	SeqRun 2	SeqRun 3	SeqRun 4	SeqRun 5
g__Pseudomonas	TRUE	TRUE	TRUE	TRUE	TRUE
c__Gammaproteobacteria	TRUE	TRUE	TRUE	TRUE	TRUE
f__Comamonadaceae	TRUE	TRUE	TRUE	TRUE	TRUE
f__Pseudomonadaceae		TRUE	TRUE	TRUE	TRUE
g__Propionibacterium			TRUE	TRUE	TRUE
f__0319-6G20			TRUE		TRUE
g__Corynebacterium				TRUE	TRUE
o__MLE1-12	TRUE	TRUE			
g__Ralstonia		TRUE			TRUE
g__Sphingomonas	TRUE				
g__Cloacibacterium	TRUE				
g__Chryseobacterium	TRUE				
g__Acinetobacter		TRUE			
o__Pseudomonadales;Other			TRUE		
g__Staphylococcus				TRUE	
g__Delftia					TRUE
f__Caulobacteraceae					TRUE
g__Herbaspirillum					TRUE

Supp. Table 2-3. Comparison of shared bacteria between negative controls and BAL samples for each sequencing run. Each column represents a sequencing run, and "TRUE" indicates that the bacterial taxa was found in at least one negative and at least 10% of the BAL samples at >0.1% abundance. Bacterial taxa are ranked by frequency across all negative control samples.



Supp. Figure 2-3. Filtering abundance thresholds tested against known bacterial mix. Mix consisted of bacterial 16S DNA from 18 unique genera and was prepared with the same protocol as BAL samples. Filtering was performed on a per sample basis, keeping genera that were present at least at the minimum threshold percent (on x-axis labels above).

Chapter 3

Investigation of lung and gut microbiota associated with a murine model of lung cancer

The work in this chapter was conducted in collaboration with Dr. Chengcheng Jin and Dr. Tyler Jacks at the Koch Institute of MIT. Mouse experiments presented in this chapter were performed and the related biological samples were collected in the Jacks lab; microbiota sequencing and analysis were performed by the author in the lab of Dr. Paul Blainey.

3.1 Abstract

The lung microbiota is altered in lung diseases such as asthma, chronic obstructive pulmonary disease, and cystic fibrosis. The microbiota may both influence disease progression and in turn be influenced by the altered lung state, but this connection is not well studied. In this work, we aim to investigate the lung and gut microbiota associated with lung cancer in collaboration with the lab of Dr. Tyler Jacks. Lung cancer is the leading cause of cancer-related deaths worldwide, with 10-15% of cancers arising in never smokers. Other factors besides smoking play a role in disease initiation and progression, with evidence showing that pulmonary infections are associated with lung carcinogenesis. This suggests a role for lung bacteria in cancer, but the biological mechanism is unclear. In addition to instances of acute infection, the lung is also host to a microbial community in the healthy state, the lung microbiota. While the gut microbiota has been shown to have a role in gastric and colorectal cancers, the role of the microbiota in

lung cancer is not understood. Here we examine the lung microbiota in a genetically engineered lung cancer mouse model using our optimized methods for lung microbiota sampling. Through collection of lung lavage fluid, optimized DNA extraction methods, and selective amplification of the 16S rRNA gene, we analyzed the lung microbiome and found that microbial diversity is reduced in tumor-bearing lungs. We discovered that tumor growth is associated with increased bacterial load in the lungs and that depletion of the microbiota through antibiotic-treatment significantly protects mice from tumor development. This study provides evidence for an association of the microbiota with lung carcinogenesis.

3.2 Introduction

As outlined in Chapter 2, the lung microbiome is a challenging system to study, but it has clear effects on human health. The lung microbiome has been associated with development of asthma [34], [35], [52], [92], disease severity in chronic obstructive pulmonary disease [30], [93], [94] and cystic fibrosis [38], [95], and proper lung structure development [36]. The presence of specific bacteria in the lung microbiota of children has been positively associated with asthma [96], and adult asthmatics have an altered lung microbiota with an increased proportion of Proteobacteria [92], [97]. Furthermore, there is evidence for a microbial community shift away from the Bacteroidetes phylum in advanced chronic obstructive pulmonary disease [30], [98], [99].

Through the development of our lung microbiome sampling methods, we were interested in studying the effect of the lung microbiota in lung cancer, an important chronic lung disease. The lab of Dr. Tyler Jacks focuses on the study of lung cancer and has developed a genetically engineered mouse model with inducible tumorigenesis. In collaboration with the Jacks Lab, we set out to examine the lung microbiota in this mouse model of lung cancer.

Lung cancer is the leading cause of death due to cancer worldwide, with a five-year survival rate of 11% [100], [101]. The most common form is non-small-cell lung cancer (85% of all lung cancers). While smoking is the main risk factor, 10-15% of lung cancers arise in never smokers [102], indicating that other factors are at play. Studies have shown that the immune system response [103], viral infections [104], [105], bacterial infections [106]–[110], and history of lung disease [111], [112] can play a role in carcinogenesis.

Specifically, there is epidemiological evidence to support a role between lung pathogens and lung cancer. Studies have demonstrated an association with viral and bacterial pulmonary infections in lung cancer patients. Prior tuberculosis infection is associated with lung cancer [108], [113]–[115], and human papillomavirus (HPV), human immunodeficiency virus (HIV), and chlamydia pneumonieae infections have all been associated with lung cancer risk [105]–[107], [116]. These associations indicate a role of lung bacteria in lung cancer progression, whether through a direct or indirect method.

There is also mounting evidence linking the commensal microbiota and cancer. Both harmful and protective effects have been found from the microbiota, mainly in gastric, oral, and colon cancer. For example, *Helicobacter pylori* infection generates a strong immune response and is associated with an increased risk of gastric cancer [117]–[119]. Additionally, *Fusobacterium nucleatum* is enriched in colorectal cancer and has been show to promote tumor growth and metastasis [28], [120], [121]. Recently, multiple studies have demonstrated the positive effect of the microbiota on cancer, primarily through priming the immune system for response to immunotherapy [122]–[126]. While epidemiological evidence can indicate an association of certain species with cancer progression, molecular and animal studies are necessary to determine the causality, as many of the studies above have demonstrated.

A multitude of studies have examined the role of microbiota in colorectal, gastric, and oral cancers, but there has been limited research in lung cancer. A number of human studies have recently examined lung cancer patients to determine whether specific bacteria are associated with increased cancer progression [127]–[129]. These studies have identified a variety of bacterial taxa significantly enriched in cancer patients versus healthy controls, but results between studies vary and are highly dependent on sampling and processing methods. Mouse models provide an ideal system to examine the role of the microbiota on lung cancer systematically as well as bypass contamination issues with human lung sampling. Through mouse models, it is possible to introduce perturbations such as antibiotics to test the direct effect of an altered microbiota and follow-up with any findings from human research. To date, no study has examined the mouse lung microbiota in a model of lung cancer.

In this work, we collaborated with the Jacks Lab to characterize the local and distal microbiota associated with lung cancer development in an autochthonous model of lung adenocarcinoma. As data from the Jacks Lab suggested depletion of the microbiota through antibiotic-treatment significantly protects mice from tumor development, we further analyzed the lung and gut microbiome upon different antibiotic treatments to explore the specific changes in the bacterial communities that may regulate tumor growth. We found that tumor growth was associated with increased bacterial load in the lung, and the lungs of tumor-bearing mice had reduced microbial diversity. This study provides evidence for an association of the microbiota with lung carcinogenesis.

3.3 Results

Methods outlined in Chapter 2 were applied to lung samples from a genetically engineered mouse model of lung adenocarcinoma. Briefly, we acquired bronchoalveolar lavage (BAL) fluid by flushing the lungs of mice with saline solution. We previously demonstrated that BAL provides superior results compared to whole lung tissue for microbiome analysis, as the amount of host mammalian DNA is far lower in BAL samples compared to tissue homogenate. We extracted DNA from BAL, selectively amplified bacterial DNA with 16S rRNA primers, and sequenced the resulting DNA library. We utilized our protocol that included optimized DNA input, 16S primers, and PCR cycling conditions for use with mouse BAL.

We hypothesized that the lung microbiome may be altered in disease, specifically in lung cancer. To test the effect of the microbiome on lung cancer, we collaborated with the lab of Tyler Jacks to utilize a genetically engineered mouse model developed in their lab. This mouse model is an autochthonous model of human non-small cell lung cancer (NSCLC). An autochthonous model is one in which the tumors arise in the mouse spontaneously or by induction, rather than transplantation. These models may more closely imitate the molecular and genetic changes that occur in human cancer [130], as tumor progression occurs slowly and evolves over time. NSCLC often has activating point mutations in *KRAS* and inactivation of the p53 pathway [131]. The Jacks Lab has established a mouse model of human NSCLC that uses intratracheal delivery of virus to activate the oncogene *KrasG12D* and induce loss of function of *p53* in lung epithelial cells [132]. This results in tumor formation 2-3 weeks after infection in $Kras^{LSL\text{-}G12D/+}$; $p53^{flox/flox}$ mice (KP mice).

3.3.1 Lung and gut microbiome in mice with lung cancer

To examine the microbial community composition in KP mice and healthy controls, we performed sequencing of the V1-2 region of the 16S rRNA gene from mouse BAL samples. In sequencing 216 mouse samples (and 29 negative controls), we identified 470 OTUs at 97% identity from a total of 8.7 million sequencing reads (see methods for filtering steps). Additionally, we sequenced fecal pellet (FP) samples from all mice using the same 16S variable region. We generated 4.4 million sequence reads and identified 420 OTUs at 97% identity in the FP samples. The most abundant phyla in fecal samples were Firmicutes and Bacteroidetes, and common bacterial families were *Lactobacillaceae, S24-7, Prevotellaceae,* and *Clostridiales* (Figure 3-1). In BAL, the top phyla were Proteobacteria and Firmicutes, and common bacterial families included *Pseudomonadaceae, Pasteurellaceae, Streptococcaceae,* and *Staphylococcaceae.* Overall, similar bacterial taxa were abundant in tumor-bearing mice (KP) and control mice. Chapter 2 of this thesis discusses BAL sequencing results and how to interpret them given bacterial reads in negative controls. See Chapter 2 for top contaminants identified and ranking of bacterial families based on their likelihood of being true mouse BAL signal. In this analysis, we present the data without deletion of any OTUs found in negative controls.

Gut microbiome samples have a distinct community profile compared to the lung microbiome, as measured by beta diversity (Figure 3-2A). Bray-Curtis and weighted UniFrac distance measures were calculated for FP and BAL samples. The Bray-Curtis dissimilarity metric is a quantitative measure of how many species are shared within a community, and the weighted UniFrac metric also measures community similarity but incorporates information on the phylogenetic distance between organisms. FP samples were more similar to each other than to BAL samples (Bray-Curtis p<0.01, weighted UniFrac p<0.01, PERMANOVA), indicating a clear distinction between the two microbiome communities. To test whether tumor-bearing mice had a distinct microbiome from healthy mice, we examined the beta diversity between KP and control mice (Figure 3-2B,C). However, we did not find any significant differences in either the lung or gut microbiota of these mice, indicating that overall microbial composition is not drastically altered in tumor-bearing mice.

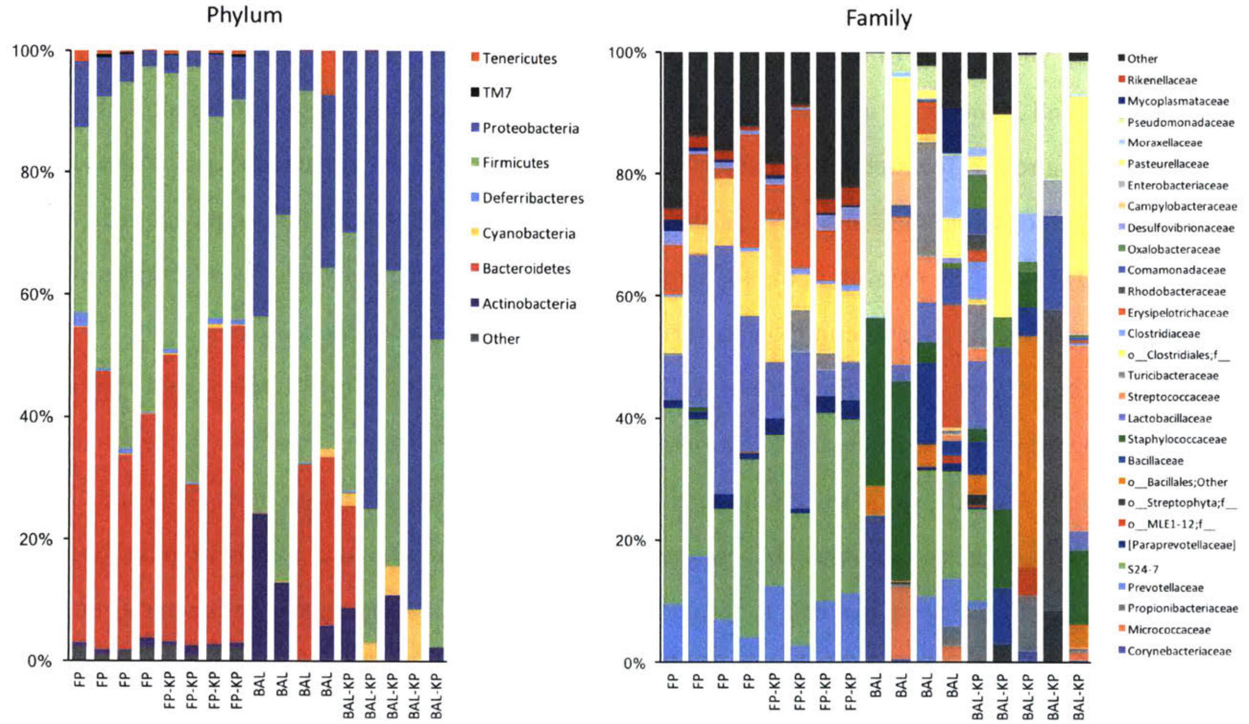


Figure 3-1. Microbial community in lung and gut in tumor-bearing versus healthy mice. Phylum (left) and family (right) level classification of gut microbiome and lung microbiome. Each bar is a summary of 2-5 mice from each experimental cohort, representing either tumor-bearing (KP) or healthy mice. FP = fecal pellet, BAL = bronchoalveolar lavage.

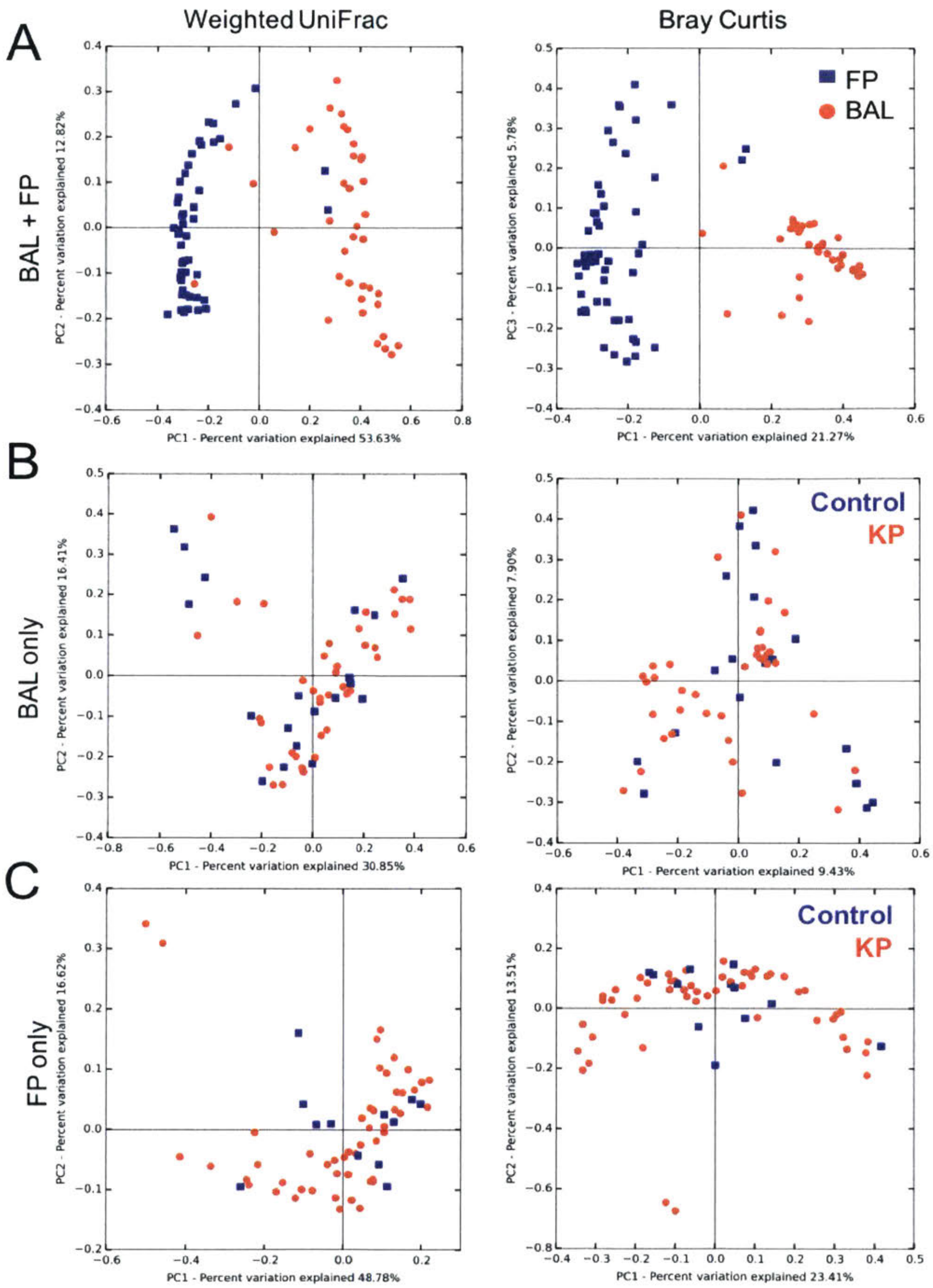


Figure 3-2. Principal components analysis (PCoA) using weighted UniFrac metric (left) and Bray-Curtis distance metric (right). A) Combined BAL and FP samples, where the lung microbiota community is significantly different from the gut microbiota community ($p<0.01$ for both Bray-Curtis and weighted UniFrac distance metrics, PERMANOVA). B) Only BAL samples (no significant difference between tumor-bearing (KP) and control mice). C) Only FP samples (no significant difference between KP and control). FP = fecal pellet, BAL = bronchoalveolar lavage.

3.3.2 Quantity of lung bacteria correlates with increased tumor burden

As previous studies have indicated that bacterial load may correlate with disease state [92], we quantified microbial DNA in mouse BAL through qPCR of the 16S rRNA gene. In work performed with the Jacks Lab, we found that tumor-bearing lungs of KP mice had higher quantities of microbial DNA compared to healthy controls (Figure 3-3A). Interestingly, we found that local microbial load in the lungs of KP mice was strongly correlated with tumor burden, while total fecal microbial load was not (Figure 3-3B,C). Of note, fecal microbial load may not be an accurate measure of total microbes in the gut but is utilized in this case as a comparison point.

3.3.3 Reduced microbial diversity in tumor-bearing lung

We measured the alpha diversity (within-sample diversity) of our samples, as prior lung microbiome analysis has demonstrated that alpha diversity is altered with disease state. We calculated the Shannon diversity index, a measure of both bacterial richness and evenness. We found that BAL samples from tumor-bearing lungs had a reduced Shannon diversity as compared to healthy controls (Figure 3-3D). We did not find a significant difference in Shannon diversity between fecal microbiome samples from tumor-bearing or control mice (Figure 3-3E). Reduced alpha diversity in the lung of KP mice indicates that fewer microbes are present and at less even distributions.

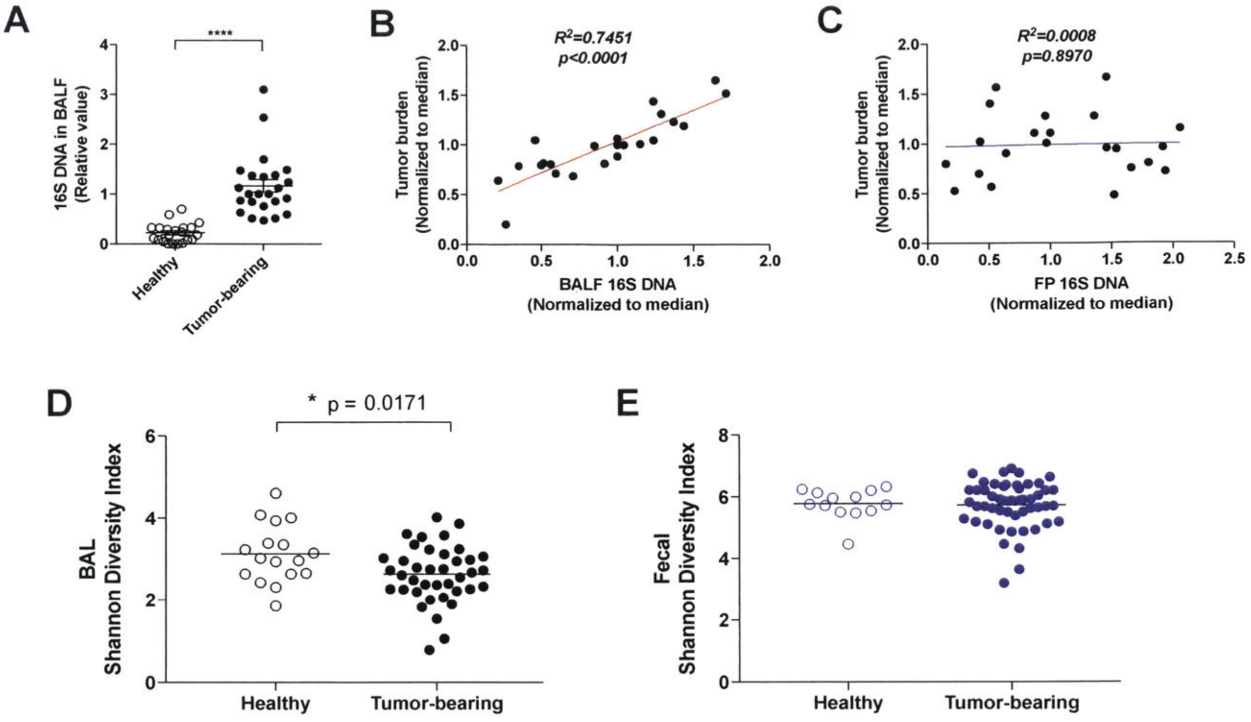


Figure 3-3. Altered microbial quantity and diversity in tumor-bearing mice. A) Tumor-bearing mice have increased lung bacterial load as measured by 16S qPCR. B) Tumor burden (normalized to median per experimental cohort) is positively correlated with microbial quantity in the lung ($p<0.001$). C) Tumor burden does not correlate with microbial quantity as measured in the fecal pellet. D) Shannon diversity is reduced in the BAL of tumor-bearing mice as compared to healthy controls. E) Shannon diversity is not significantly different in the fecal microbiome. (Data from A,B,C produced by Dr. Chengcheng Jin)

3.3.4 Specific bacterial taxa absent in tumor-bearing mice

To examine if specific bacteria were abundant or diminished in tumor-bearing mice, we performed testing with the nonparametric Mann-Whitney U test to compare ranks of bacteria in tumor-bearing versus healthy mice (relative abundance). Bonferroni correction was used due to multiple hypothesis testing. We found one bacterial taxa significantly different between the two groups, the genus *Prevotella* (Figure 3-4). No other OTU or bacterial taxa was significantly different. In the gut microbiota, we found a number of OTUs and taxa that were different, most significantly the genera *Corynebacterium* and *Staphylococcus* (Figure 3-4). All of these bacterial taxa were enriched in healthy mice but absent in tumor-bearing mice. As we did not identify any bacteria significantly abundant only in tumor-bearing mice versus controls, this suggests that no bacteria were significantly abundant in a majority of tumor-bearing mice.

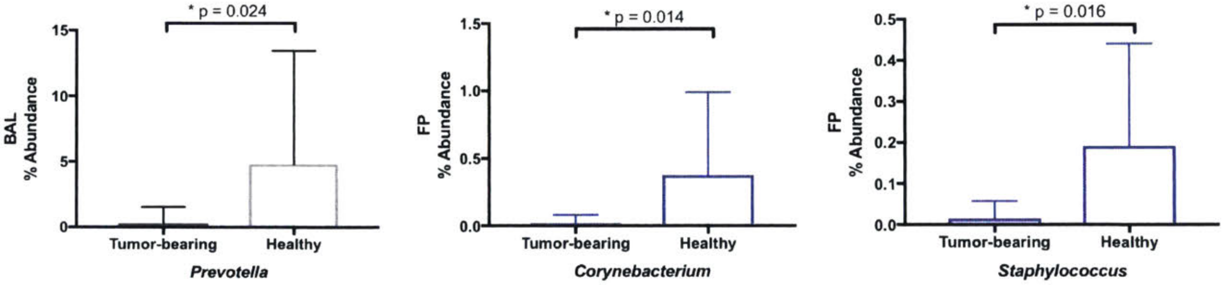


Figure 3-4. Examination of bacteria differentially abundant in tumor-bearing mice. A) The genus *Prevotella* was significantly more abundant in the lung microbiome of healthy mice (p=0.024). B) The genera *Corynebacterium* and C) *Staphylococcus* were more abundance in gut microbiota of healthy compared to tumor-bearing mice. All p values are calculated with the Mann-Whitney U test with Bonferroni correction.

3.3.5 Commensal microbiota promotes lung cancer tumor progression

To test the significance of commensal microbiota in lung cancer, the Jacks Lab performed experiments to perturb the microbiota through antibiotic treatment. Specific-pathogen-free (SPF) KP mice were treated with a cocktail of antibiotics at 6.5 weeks post-infection (initiation of tumor through viral administration). A mixture of four antibiotics (4abx) was put in the drinking water: ampicillin (1g/L), neomycin (1g/L), metronidazole (1g/L) and vancomycin (500 mg/L). These antibiotics have diverse bacterial targets and a range of absorption capacity from the gut into the bloodstream (Table 3-1). This oral administration of antibiotics resulted in a striking suppression of lung cancer progression. Tumor burden and tumor grade were significantly reduced in 4Abx-treated mice compared to age-matched controls, with tumor burden diminished by roughly 30% (Figure 3-5). Additionally, treatment with a single antibiotic, metronidazole (targeting anaerobic bacteria), was also effective in suppressing lung tumor growth. Evidence from work in the Jacks Lab indicated that antibiotics did not alter the rate of tumor cell proliferation, ruling out the potential that direct cytotoxicity from antibiotics caused the tumor phenotype. These findings provide evidence that the microbiota has a role in promoting tumor progression in this model of lung adenocarcinoma.

Antibiotic	Spectrum	Absorption through gut
Ampicillin	Gram+, Gram-	30-55%
Metronidazole	Anaerobic	> 80%
Neomycin	Gram-	3%
Vancomycin	Gram+	< 5%

Table 3-1. Antibiotics administered to tumor-bearing mice. Gram negative, gram positive, and/or anaerobic bacteria are targets of each antibiotic, and antibiotics have a range of absorption through the gut into the bloodstream.

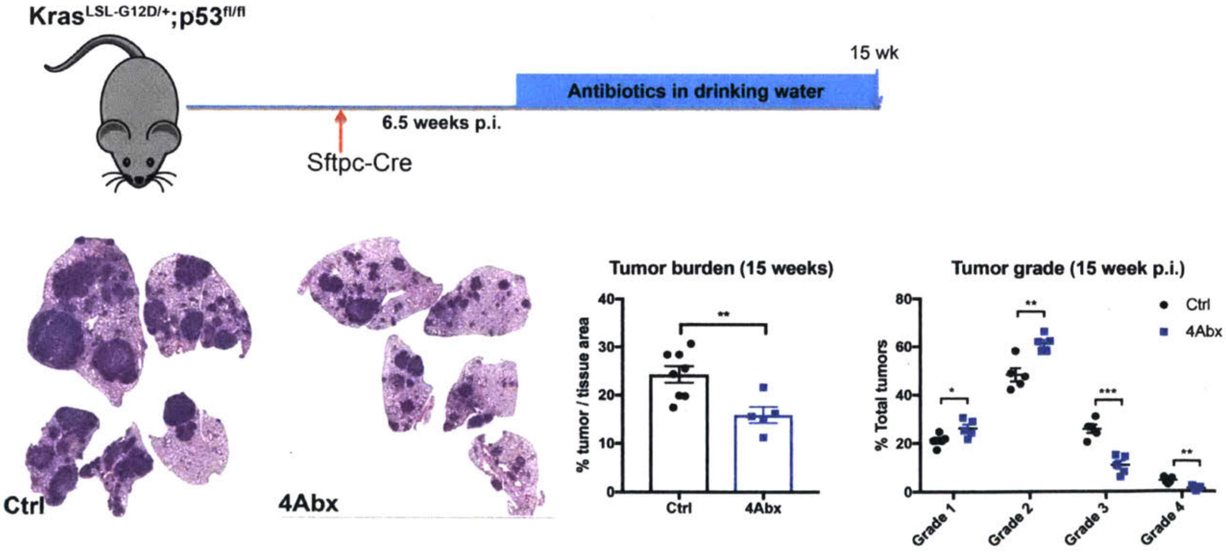


Figure 3-5. Commensal microbiota promotes tumor initiation and progression of lung adenocarcinoma. KP mice were treated with a cocktail of four antibiotics (4abx) administered via the drinking water. Antibiotic-treated mice show reduced tumor burden and fewer high-grade tumors. (Data and figure produced by Dr. Chengcheng Jin)

3.3.6 Tumor-bearing mice have reduced microbial diversity with antibiotic treatment

As antibiotics had a clear impact on tumor progression, we set out to examine if there were specific changes to the microbial composition that might have a role in this tumor growth alteration. First, we compared alpha diversity in the control KP mice and antibiotic-treated KP mice. Figure 3-6A and B demonstrates that lung microbiota was significantly reduced in antibiotic-treated mice, and reduced diversity was also found in gut microbiota. In general, antibiotics had a smaller effect on the lung microbiota diversity as compared to the gut microbiota diversity. Multiple antibiotic treatment combinations were used: 4abx (ampicillin, neomycin, metronidazole, and vancomycin), vancomycin/neomycin only, and metronidazole only. The Shannon diversity was calculated for each antibiotic treatment (Figure 3-6C,D). While neomycin and vancomycin are poorly absorbed through the intestine, metronidazole has a much higher absorption (80%). Therefore, we would expect metronidazole to be absorbed systemically and likely reach the lung environment. We might expect to see a lower Shannon diversity in the metronidazole treatment as compared to the vancomycin/neomycin treatment in the lung microbiota, but this is not the case. We do find that as the number of antibiotics increase, the gut microbiota correspondingly decreases in diversity. This is expected as the antibiotics are administered orally and have a direct local effect on the gut microbiota.

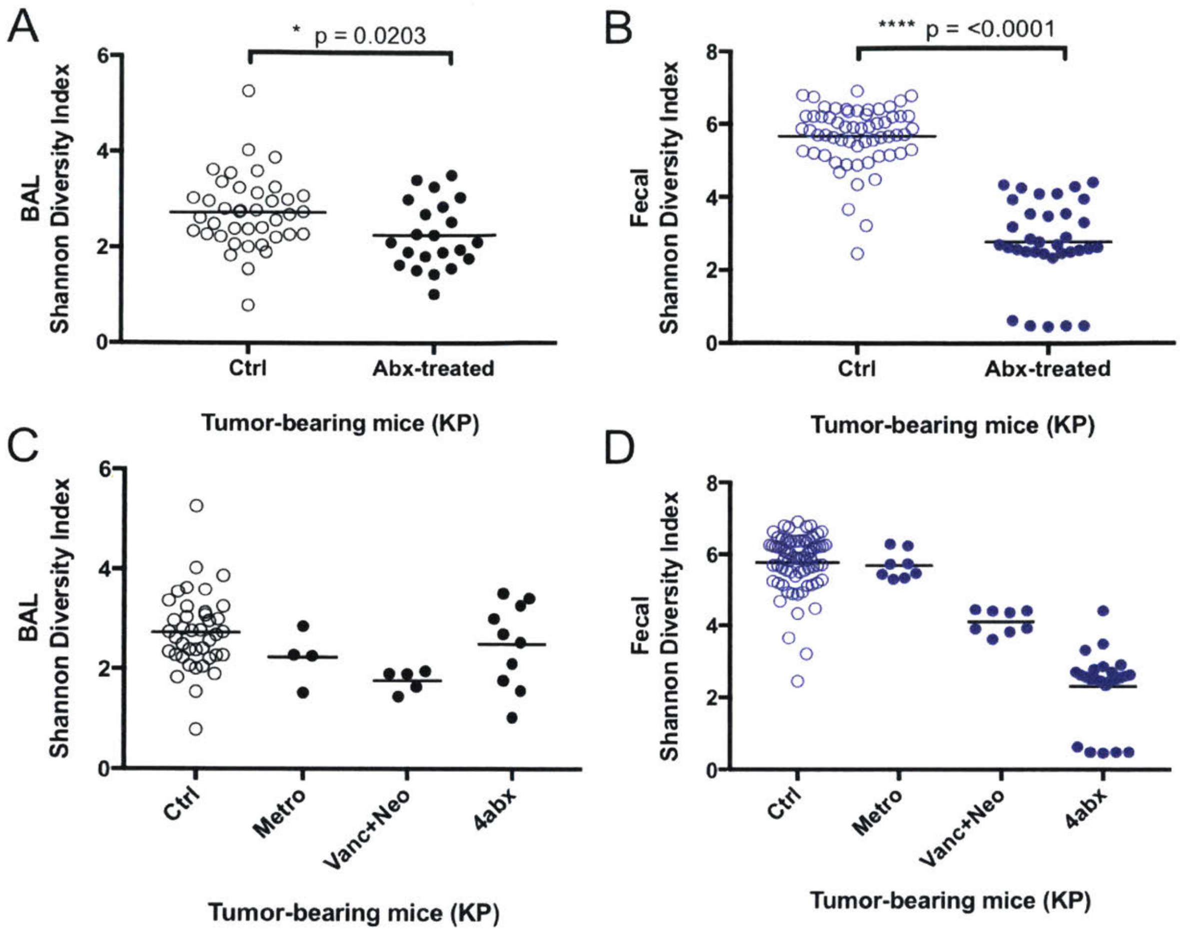


Figure 3-6. Alpha diversity in lung and gut microbiome of tumor-bearing mice. Shannon diversity index of BAL samples (A) and FP samples (B) is significantly reduced in antibiotic-treated KP mice compared to control (no antibiotics) KP mice. Shannon diversity index per each cohort of antibiotic treated mice in the BAL samples (C) and fecal pellet samples (D). Significance is determined with the two-tailed Student's t-test.

3.3.7 Alpha diversity does not correlate strongly with measured tumor burden

As antibiotics reduced the tumor burden in KP mice, there was a range of tumor burden across the KP cohort. Therefore, we tested whether alpha diversity was correlated with increased tumor burden in these mice. We found that there is a slight positive correlation between Shannon diversity and tumor burden: the gut microbiota diversity had a significant linear correlation while the lung microbiota correlation was not significant. For both cases, the explained variance from the linear fit was low. This indicates that although tumor-bearing lungs have overall lower alpha diversity, there was not a strong correlation within the subset of tumor-bearing mice.

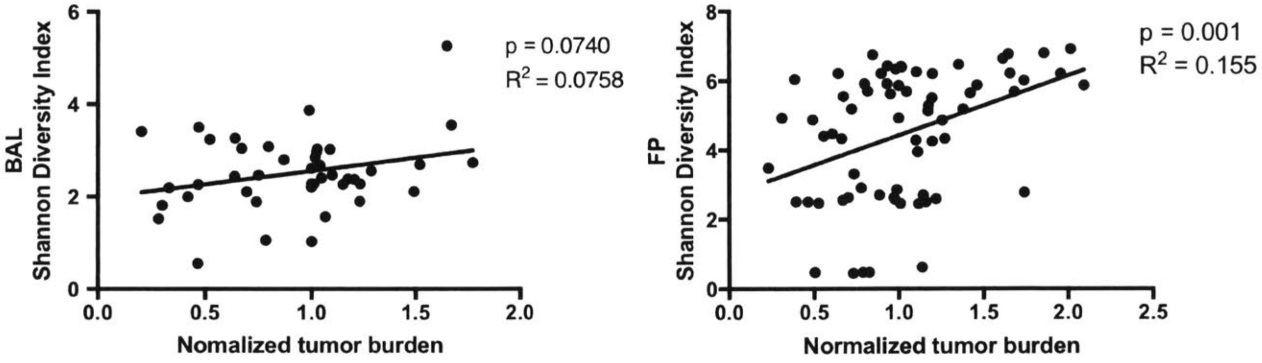


Figure 3-7. Tumor burden correlation testing with alpha diversity. Lung microbiome diversity from BAL samples (left) did not significantly correlate with normalized tumor burden. Gut microbiome diversity (right) had a significant ($p>0.05$) linear correlation but low R^2 value. Tumor burden was normalized for each sample by dividing by the median tumor burden per experimental cohort (as time of sample collection varied slightly between cohorts).

3.3.8 Tumor burden correlates with abundance of specific gut microbes

We tested whether the particular abundance of a bacteria was correlated with tumor burden using the non-parametric Spearman rank-order correlation. No bacteria were significantly correlated in the lung microbiome (using p<0.05, Bonferroni correction). This may be due the more variable lung microbiome and fewer shared bacterial taxa across samples. In the gut, the phylum Bacteroidetes, the class Clostridia, and the genus *Bacteroides* were correlated with increased tumor burden (Figure 3-8). This result was from the combined cohort of all tumor-bearing mice, control and antibiotic-treated. Further testing within cohorts can identify whether different antibiotic treatments had different microbial effects. The identification of gut bacteria correlated with tumor burden suggests that there may be cross-talk between the lung and gut during the diseased state of lung carcinogenesis. Although the overall bacterial load in the gut did not appear to be correlated with tumor burden, specific microbial composition may still have a potential tumor-regulating effect.

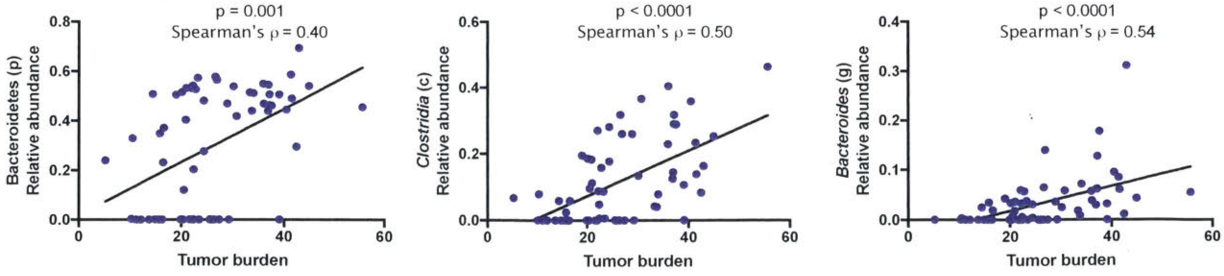


Figure 3-8. Correlation testing of specific bacterial abundance with tumor burden. Three taxa of bacteria were significantly correlated with tumor burden (Spearman rank-order correlation test for p<0.05).

3.3.9 Antibiotics have effect on gut microbiota and minimal observed effect on lung

A cocktail of four antibiotics (4abx) were administered orally to healthy and tumor-bearing mice, as described above, and this treatment resulted in decreased tumor progression in KP mice. These antibiotics have a range of bacterial targets and absorption capacity through the intestine. To determine if there were specific bacterial changes that might correspond to altered tumor progression, we examined the lung and gut microbiota. Figure 3-9 shows the phylum level summary of microbiota from untreated and 4abx-treated samples. The BAL samples are heterogeneous and do not have a distinct pattern with or without 4abx treatment, while the fecal microbiota samples show a very clear perturbation. The lung microbiota did not have any significantly altered bacterial taxa after multiple-hypothesis testing correction (Figure 3-9 shows two taxa that are altered but not significantly different with $p<0.05$, Bonferroni correction Mann-Whitney U test). While we did not identify a specific taxon or OTU that was significantly different in the lung microbiota, this does not rule out the potential that bacteria were altered in categories we did not measure, such as specific phenotypic traits or gene content or gene expression. Furthermore, as the lung microbiota is more variable than the gut microbiota, larger sample sizes in the future would better enable detection of changes. In the gut microbiota, there were many significantly altered phyla between control and 4abx treatment. Antibiotic treated samples had reduced Bacteroidetes and Firmicutes, and increased Tenericutes and Proteobacteria.

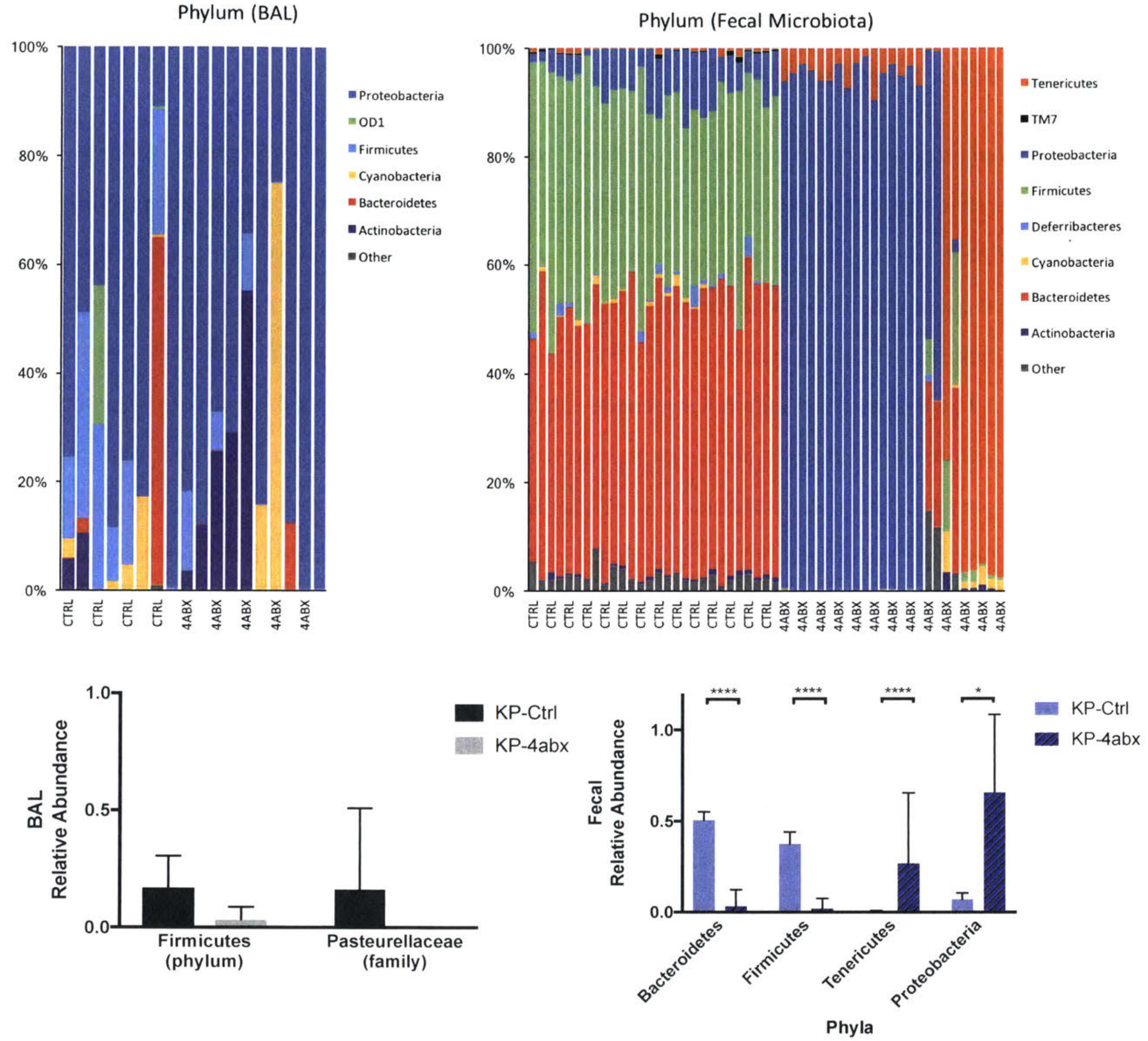


Figure 3-9. Effect on lung and gut microbial community with antibiotic treatment. A cocktail of four antibiotics (4abx) was administered orally to KP mice (control received no treatment). Top graphs show the phylum level taxonomy for BAL (left) and FP (right) samples. Bottom graphs show bacteria that are altered in 4abx treatment compared to control. BAL (left) had altered taxa but they were not statistically significant ($p<0.05$, Mann Whitney U test with Bonferroni correction). FP (right) had multiple phyla that were significantly altered with 4abx treatment.

3.4 Discussion

Mounting evidence has demonstrated a link between the lung microbiota and chronic lung disease (asthma, cystic fibrosis, chronic obstructive pulmonary disease). While studies continue to tease apart the cause and effect relationship between lung microbiota and lung disease progression, it is informative to use animal models to investigate specific disease states. We applied our methods for sampling the mouse lung microbiome to examine the role of microbiota in lung cancer.

The role of the commensal microbiota and cancer has been clearly established in certain cases. Studies have shown a harmful effect with specific bacteria tied to tumor progression (*Fusobacterium nucleatum, Helicobacter pylori*), while other studies have demonstrated that specific bacteria are necessary for response to cancer immunotherapy treatments. Most studies have focused on the gut microbiome, as it has the largest quantity of bacteria and is easily sampled. However, tumor tissue has also been investigated, and bacteria were found in breast cancer tumor samples and pancreatic tumor samples [29], [133]–[135]. Both studies examining pancreatic tumor found an increased abundance of bacteria in tumor samples compared to healthy controls. These studies suggest that the microbiota has a role in either promoting local inflammation and oncogenesis or priming an anti-tumor response. While the role of the microbiota in lung cancer is not well understood, our study aimed to investigate the effect of the lung and gut microbiota in a lung cancer mouse model.

Previous work on the lung cancer microbiota has found evidence for specific bacterial associations in tumor tissue versus healthy tissue, [127]–[129]. However, bacteria unique to each study were identified as significantly associated in lung cancer patients, so no common bacteria were identified and it is unclear if results are generalizable. Furthermore, sample types range from BAL to tissue collection to sputum, complicating comparisons. A significant similarity from these studies, however, was the finding that lung microbial diversity was reduced in patients with lung cancer. In our study using a mouse model of lung cancer, we found similar results: lung microbial diversity is reduced in tumor-bearing mice compared to controls. We did not identify any significantly enriched bacterial taxa between tumor-bearing and healthy mice. Furthermore, we identified that tumor growth is associated with increased lung bacterial load, but not gut bacterial load.

3.4.1 Effect of lung disease on the respiratory environment

To put our results in context, it is necessary to take into account the local lung environment and how it changes with disease. First, the lung has a network of immune cells to maintain homeostasis, as the lungs are constantly challenged with new inhaled particles or bacteria [136]. The immune system can be strongly shaped by the local microbiota, and studies have identified links with Th17 responses [137], macrophage phenotypes [138], and inflammatory cytokine profiles [139] in the lung. Recent work from Jin et al. demonstrated that tissue-resident γδ T cells are activated by lung microbiota and in turn produce IL-17 to promote inflammation and further tumor growth (C. Jin, T. Jacks, et al., work in preparation). This indicates that the immune system is shaped by the local microbiota in the lung.

The presence of tumor lesions in the lung induces a range of perturbations to the normal lung environment. First, the immune system is altered with recruitment of inflammatory cells (such as neutrophils and macrophages) and production of pro-inflammatory cytokines near the site of tumorigenesis [140]. Second, physiological changes occur in the lung due to tumor growth. As lung tumors grow in size, they may diminish or block airway passages, and this in turn can impair mucociliary clearance. As this is the main method for microbial elimination, this reduced elimination rate may explain the increased bacterial load we found in the tumor bearing lungs. Furthermore, injury and inflammation is associated with increased mucus production in the lung, providing an increased nutrient supply to bacteria in a normally sparse environment [141]. Dickson et al. propose that as lung disease worsens, the lung microbiome composition is determined more by reproductive rates of the bacteria and local growth conditions rather than immigration and elimination in the lung. Increased bacterial burden has been observed in patients with severe destructive lung disease [142], and pneumonia provides an example of how overgrowth of bacteria is promoted due to host inflammatory defense mechanisms. A positive feedback loop is generated that supports the dominance of a specific pathogen in the lung microbiota [143].

These factors may explain why we found an increased bacterial load in tumor-bearing lungs. Our result was similar to the studies that found increased bacterial load in pancreatic tumor samples. Furthermore, our finding that microbial diversity was decreased in tumor-bearing lungs is supported by the notion that the altered lung environment selects for growth of specific bacteria. Selective growth of bacteria would reduce community diversity, and hindrance of mucociliary clearance would diminish the removal of growing lung microbes. On the other hand, we did not find a signature of one or a few specific microbes that were common across tumor-bearing mice. This indicates that the lung environment is not highly selective, the residence time of microbes is very short, and/or the environment is highly dynamic. Work from others support this notion that the lung is not stably colonized but instead is in a constant state of flux [144].

In comparison with the lung microbiota, the gut microbiota did not show a change in alpha diversity or total microbial load with lung cancer. The lung microbiota is more dynamic and transient than in the gut due to its lower bacterial burden and constant microbial immigration and elimination [145]–[147]. Also, the gut microbiota did not experience a local change in environment as found in the lung with tumor growth. However, there is likely cross-talk between the gut-lung microbiota mediated by the immune system and systemic microbial metabolites [148], [149], so an effect on one environment is likely to impact the other.

In examining the effect of antibiotics in the lung microbiota, we found no clear signal showing consistent microbial changes across samples. Again, this supports the idea that the lung microbiota is a dynamic environment with many forces as play, both for microbial elimination and immigration. While we did test a variety of antibiotics, we cannot rule out that the antibiotic type or concentration was not effective in altering lung microbes. There are technical challenges associated with using antibiotics to specifically deplete the lung microbiota. While oral administration is the easiest delivery method, the antibiotics first pass through the gut and may or may not be absorbed through the intestine. We attempted to selectively deplete the lung microbiota without perturbing the gut microbiota through aerosolized antibiotics, but this was very challenging and the amount of inhaled antibiotic was hard to control. Furthermore, we used different types of antibiotics to test whether high or low absorption capacity through the gut would influence the effect on the lung. While some antibiotics like vancomycin and neomycin are not expected to be absorbed through the gut, small amounts could reach the lung through microaspiration when ingesting the water/antibiotic mixture. Therefore, there are challenges with using antibiotics to selectively deplete either the lung or gut microbiota. In this work, we only describe in-depth analysis with the cocktail of four antibiotics, as this experimental cohort had a large enough sample size. Due to sample processing issues, the metronidazole-treated cohort did not have enough samples for robust analysis. However, future work will examine metronidazole-treated mice to examine if there are changes in the abundance of anaerobes or other specific bacteria.

3.4.2 Directionality of causation

Does increased bacterial burden and altered community drive the progress of lung oncogenesis? Or is it just a consequence of altered growth environment in the lungs? We find that depletion of microbiota in SPF antibiotic-treated mice protects against tumorigenesis, so this indicates that the microbiota alone plays a role in driving the progression of disease. However, it is unclear what exact community members in the commensal microbiota specifically drive the disease progress. Jin et al. (in submission) demonstrated that the presence of microbiota activates a subset of T cells in the lung which in turn activate the immune system to increase tumorigenesis. This establishes a clear role of the microbiota in influencing the immune system.

Dickson et al. present a model of "dysbiosis-inflammation cycle" in lung disease that takes into account a bidirectional relationship with lung microbiome and host response [141]. Any source of inflammation (such as tumor growth) may provoke host responses to rapidly alter the lung environment. Increased mucus production, generation of inflammatory cytokines, targeting of specific bacteria, and altered oxygen concentration may all play a role in favoring growth of specific bacteria or a total increase in bacterial load. This shift in community can then send signals back to the immune system that might provoke further inflammation, potentially through a higher presence of pathogen-associated molecular patterns and microbial metabolites that activate the immune system [146].

Although we did not find specific members of the lung microbial community that were associated with increased tumor growth, our results suggest that the lung microbiota is altered in tumor-bearing lungs and correlates with increased disease progression. We did find a reduced relative abundance of *Prevotella*, and further analysis will be performed to examine changes in absolute abundance. As we describe in Chapter 2, the lung microbial community is variable across healthy mice, indicating that the community is dynamic and constantly changing. This property makes it difficult to identify a specific bacterial taxon that correlates with lung cancer progression, as the signal is likely distributed across many microbial community members.

Future work can further investigate the "noisy" lung microbiome signal in lung cancer and apply computational methods to examine correlation networks that may exist in the current data. The OTUs and taxonomic categories derived from 16S sequencing may not provide the proper categories for comparison, as microbial gene content, gene expression, secreted metabolites, or other functional categories may have a more important role in discriminating which bacteria have a direct effect in lung carcinogenesis. Furthermore, comparison of phenotypic properties of the lung microbiota may shed light on the different roles of the community members.

In order to follow-up on the results from this study, it would be ideal to perform culture or bacterial-specific 16S PCR to identify specific microbes in the lung. We had difficulty with any success in culture of mouse lung microbes, but this would verify sequencing results. Furthermore, an extension of this study may include working with mice in non-SPF conditions, as done by Yun et al., where they had greater diversity of microbes and greater success of culture [36]. Another extension would be to sample excised tumor tissue along with healthy adjacent tissue in the tumor-bearing lung, which would provide insight into whether there are bacteria unique to each site. In summary, we found that commensal microbiota promotes tumor progression in lung cancer, suggesting a role of microbiota in lung carcinogenesis.

3.5 Methods

3.5.1 Animal work

Mouse facility work is detailed in the Methods of Chapter 2.

3.5.2 Sample collection

Both bronchoalveolar lavage (BAL) and fecal pellet (FP) samples were collected from each mouse. Additionally, one blank control sample was collected along with each mouse cohort to provide a negative control for BAL collection, DNA extraction, and sequencing preparation. Sample collection is described in further detail in Chapter 2 Methods.

3.5.3 DNA extraction and sequencing preparation

The phenol-chloroform extraction method outlined in Chapter 2 Methods was used for all BAL and FP samples. The V1-2 variable region of the 16S rRNA gene was amplified for Illumina sequencing using a two back-to-back PCR reactions, as described in Chapter 2. FP samples were prepared in the same way as BAL samples, except only 50 ng of DNA was used for the starting material in the first PCR reaction and only 10 cycles were performed for the second PCR reaction. This was due to the fact that the FP samples had a much higher concentration of microbial DNA and did not require as many amplification cycles.

3.5.4 qPCR to quantify microbial DNA

In order to measure the amount of microbial DNA, qPCR was performed with the primers 27F (AGAGTTTGATCMTGGCTCAG) and 338R (TGCTGCCTCCCGTAGGAGT) that target the V1-2 region of the 16S rRNA gene. 100ng of DNA was added to 10μL Kapa SYBR Fast 2X master mix (Kapa Biosystems), 1μL 10uM forward primer, 1μL 10uM reverse primer, 1μL Eva dye, 0.2μL Rox reference dye, and PCR-clean water to bring the total reaction volume up to 20μL. The reaction was then run at 95C for 3 minutes; 45 cycles of 95C for 15 seconds, 61C for 60 seconds; and 72C for 1 minute. *E. coli* genomic DNA was used as a standard, with samples at 1, 0.1, 0.01, and 0.001 ng per reaction used to create a standard curve.

3.5.5 Sequence analysis

Sequencing quality filtering and OTU assignment is described in Chapter 2 Methods. Briefly, OTUs were clustered at 97% identity using QIIME *pick_open_reference_otus.py* with the Greengenes 13.8 reference database. OTUs that did not appear in at least 3 samples and have at least 50 sequence counts were removed. On average, BAL samples had 50,000 sequencing reads and FP samples had 10,000 sequencing reads. A minimum threshold of 500 reads was used to retain samples after filtering. OTUs were assigned taxonomy to the deepest taxonomic level with a minimum 90% similarity from the Greengenes database, and taxonomic classifications ranged from phylum down to genus level.

To calculate diversity metrics, the QIIME pipeline was used. For alpha diversity, we calculated the Shannon index of the 97% identity OTUs using *alph_diversity.py*, with samples rarefied to 1000 reads. Beta diversity measurements were calculated using *beta_diversity.py* to determine the Bray-Curtis, weighted UniFrac, and unweighted UniFrac metrics. PCoA plots were made using *make_2d_plots.py*.

3.5.6 Statistical analysis

We utilized the Student t-test in GraphPad Prism 7 to compare Shannon diversity indices (with significance for $p<0.05$). To test if bacteria were significantly altered in abundance between cohorts, the non-parametric Mann-Whitney U test was used to compare the ranks of bacterial abundance per sample, with a Bonferroni correction for multiple hypothesis testing (using QIIME *group_significance.py*). Testing was done at the OTU level as well as each taxonomic level with relative abundance counts, with p values corrected according to the number of categories. To test if specific bacterial abundance correlated with tumor burden, the non-parametric Spearman rank-order correlation was used with QIIME *observation_metadata_correlation.py,* as bacterial abundance has a non-normal distribution. Normalized read counts were used (relative abundance). Bootstrapping with 1000 permutations was performed to calculate p values, and a Bonferroni correction was used due to multiple hypothesis testing. Linear correlation analysis was performed in GraphPad Prism to test if alpha diversity correlated with tumor burden, and p values derive from testing the null hypothesis that the overall slope is zero.

To test for differences in community composition between sample groups, we performed Permutational Multivariate Analysis of Variance (PERMANOVA) based on the Bray-Curtis and weighted UniFrac distance matrices using QIIME *compare_categories.py.* Statistical significance between groups was determined using 1000 permutations of sample labels.

Chapter 4

Microfluidic sample preparation enables efficient large-scale bacterial whole genome sequencing

Work in this chapter was performed in collaboration with Mohamad Sater and Yonatan Grad.

4.1 Abstract

High-throughput and low-cost whole genome sequencing (WGS) of bacterial pathogens has the potential to transform clinical microbiology and infection control through high-volume surveillance of patient bacterial specimens. However, sample preparation of bacteria for sequencing is more resource intensive, costly, and logistically complex than DNA sequencing, providing a serious barrier for large-scale studies (100s – 1000s of isolates). In this work, we implemented an improved operating protocol for a fully integrated cells-to-sequence library workflow that increased the throughput of our custom microfluidic platform by an order of magnitude. The new procedure produced DNA libraries at a rate of 1000 samples per 7 days with a 100-fold reduction in cost, taking advantage of reduced reaction volume, integrated sample preparation, and automation. The microfluidic sample preparation produced DNA libraries that were reproducible and of comparable quality to libraries from a standard benchtop preparation method. We applied this method to prepare and sequence a total of 4000 isolates, of which 3000 are methicillin-resistant *Staphylococcus aureus* (MRSA) collected from study subjects in a clinical trial of the impact of an MRSA decolonization protocol. Using WGS data, we identified 45 different MRSA sequence types and determined that samples were resistant to an average of 4.5 antibiotics (range 1-9). Our study

demonstrates that this microfluidic technology can facilitate large-scale bacterial WGS to support clinical and epidemiological studies.

4.2 Introduction

Whole genome sequencing provides the highest-resolution comparison between bacteria, detecting differences down to the level of a single nucleotide. WGS can detect the presence of antibiotic resistance genes and enable construction of pathogen transmission networks to aid with epidemiological studies. Clinicians can use information from WGS to decide on treatment options and monitor the spread of pathogens, with a transformative effect in clinical microbiology [150]. With the rising incidence of antibiotic resistant bacteria including methicillin-resistant *staphylococcus aureus* (MRSA), detection and tracking of pathogens is an important issue in human health [151], [152]. The decreasing cost of sequencing (now on the order of $1 per bacterial genome) has enabled large-scale epidemiological studies, and there has been a 100-fold increase in the average number of bacterial genomes per study over the past 10 years. Large-scale studies have powered the earlier detection of outbreaks [41], accurate characterization of transmission networks [42], [43], detection of the timing and mechanism of antibiotic resistance acquisition [44], and bacterial genome-wide association studies [45]–[47]. With the capability to sequence at higher throughput, large-scale bacterial sequencing is poised to help track the spread of pathogens and evaluate treatment methods in real-time. Already, researchers are examining MRSA population structure across nations and integrating WGS data into routine pathogen surveillance [153].

Despite the promise of WGS to improve clinical care, routine adoption of WGS in clinical microbiology is limited not by sequencing cost but rather by sample preparation. To sequence bacterial DNA, researchers must lyse cells, extract and purify DNA, normalize DNA input, fragment and tag DNA with sequencing adapters, purify the resulting DNA library, amplify the library with barcoded primers, pool samples, and load onto a sequencing instrument. This laborious process places high demands on time, labor, cost, and expertise, resulting in at least a ten-fold higher cost compared to sequencing. Multiple reagent kits need to be used during the protocol and throughput is limited without expensive robotic equipment. Current library preparation methods require one million cells and perform reactions in volumes of 5-50 μL. To overcome the challenges of reagent cost, others have developed reduced-volume methods on the benchtop [154]–[156] and with acoustic liquid dispensing systems [157].

However, these methods still require multiple steps of sample handling, consumables like pipette tips, extensive labor, and potentially expensive liquid handling equipment. In all cases, genomic DNA must be carefully normalized into the reaction for sequencing preparation. No solution exists that provides end-to-end sequencing library preparation on a single platform, starting with whole cells.

To overcome the sequencing preparation challenge, we developed a streamlined protocol that automates the entire process in one device at small volume. We developed this protocol based on the microfluidic device designed by Kim et al., which converts normalized DNA into sequencing libraries in nanoliter-scale reactions [158]. The method by Kim et al. produced about 400 bacterial libraries with one-time-use devices that process 16 samples per day (for a total of 25 devices). This method drastically reduced the cost per sample while retaining data quality. However, this protocol has limitations in throughput for large sample sizes, since only 16 samples can be processed per day and new devices are required for each run.

Our new protocol enables sample preparation and sequencing of thousands of bacterial isolates by a single researcher in the lab through automation, device reusability, and robust/reliable workflow. Our protocol increases throughput by a factor of ten compared to Kim et al., enabling sequencing preparation of 1000 samples in 7 days. Cells are loaded directly into the device, allowing for automated DNA extraction and normalization in the device. We apply this method to sequence 4000 bacteria isolates of diverse GC content, of which 3000 are MRSA isolates from a clinical trial testing decolonization. Through WGS, we identify 45 different MRSA sequence types (majority ST5 and ST8) and screened for resistance to 12 different antibiotics. This application demonstrates the feasibility of routine epidemiological studies in the clinic using this methodology.

4.3 Results

4.3.1 Microfluidic device protocol

We developed a high-throughput, reproducible, and translatable protocol for genome sequencing preparation building on an existing microfluidics technology [158]. Microfluidics enable the manipulation of small volumes (nanoliters) with high precision and automation. Preparation of bacterial samples for whole genome sequencing requires cell lysis, DNA extraction, DNA fragmentation and tagging with sequencing adapters, DNA library purification, and barcoding. Microfluidics enable protocol automation and reduced reaction volume, removing technical labor and cost barriers to high-throughput sample preparation. We applied the microfluidic device design demonstrated in Kim et al. in a new protocol that allowed for sequencing preparation of 4000 bacterial isolates, with a rate of approximately 1000 isolates per 7 days (Figure 4-1).

The microfluidic device used in this protocol has a set of 36 reactors, each with a 40 nL reaction volume. Four devices are run in parallel to achieve a throughput of 144 samples per 7.5 hours (4 hours of hands-on time). Bacterial cells are input directly to the device, and the library preparation steps proceed in an automated fashion, requiring only input of enzymes by the user. In comparison, Kim et al. used a 16-reactor device for each run, requiring approximately 8 hours of hands-on time. In our new method, we were able to achieve an order of magnitude more throughput with decreased hands-on time (Figure 4-1A). We also re-used devices for each run, while Kim used a new device each time, requiring over 25 devices for 400 samples.

The details of the device design and basic operating procedure are previously described [158]. Briefly, inside each of the reactor chambers, cells were lysed with an enzyme mixture, DNA was purified with SPRI beads, Nextera enzyme was added to fragment and tag the DNA with sequencing adapters, and the DNA library was eluted from the device (Figure 4-1A). Subsequently, DNA libraries were barcoded and amplified in 96-well plates and then loaded onto a sequencer.

4.3.2 Reuse and reliability of microfluidic devices

As fabrication of microfluidic devices can be time-consuming and technically challenging, we developed a device cleaning procedure that enabled reuse by eliminating cross-contamination. For each run, one negative control (no cells) and one positive control (different cell type/distant species) were loaded into the device, along with 34 test samples. As shown in Figure 4-1B, *Staphylococcus aureus* samples were loaded next to a negative control (blank reactor without cells) and a different cell sample (*Escherichia coli*). After preparing DNA libraries, the device was cleaned with a potassium hydroxide solution which dissolves the residual DNA due to the high pH. In the subsequent device run, the *E. coli* and blank sample are shifted to new reactors. Placing the positive control (*E. coli*) and negative control (no cells) in different reactors for the next run enabled us to test if there was any cross-contamination between reactors in one run or within the same reactor of back-to-back runs. We tested for cross-contamination by sequencing the barcoded samples and searching for contaminating DNA. We found that blank samples (no cells) had no detectable amount of *S. aureus* or *E. coli* DNA, and there was no cross-contamination between *E. coli* and *S. aureus* reactors. We used this device setup with positive and negative controls in all of our sample processing, facilitating sequence-based evaluation of contamination across runs.

In addition to device reuse, device and protocol reliability is critically important. To process thousands of samples, the protocol on each device needs to be repeated over a hundred times. Previously, new devices were fabricated each time and the protocol had <75% success rate. Failures could be due to human error or device malfunction. We automated the entire protocol with a custom Matlab script and optimized device set-up, achieving a success rate >95%. As shown in Figure 4-1C, four devices were utilized over 40 times each, with no more than two run failures per device. Overall, four devices processed 90% of the samples, and five devices processed all samples. One device had to be replaced during our processing due to a clogged valve (no devices failed due to mechanical failure of the valves). This high reuse capacity means that very little microfabrication was necessary to process thousands of samples; devices can be used repeatedly with proper controls.

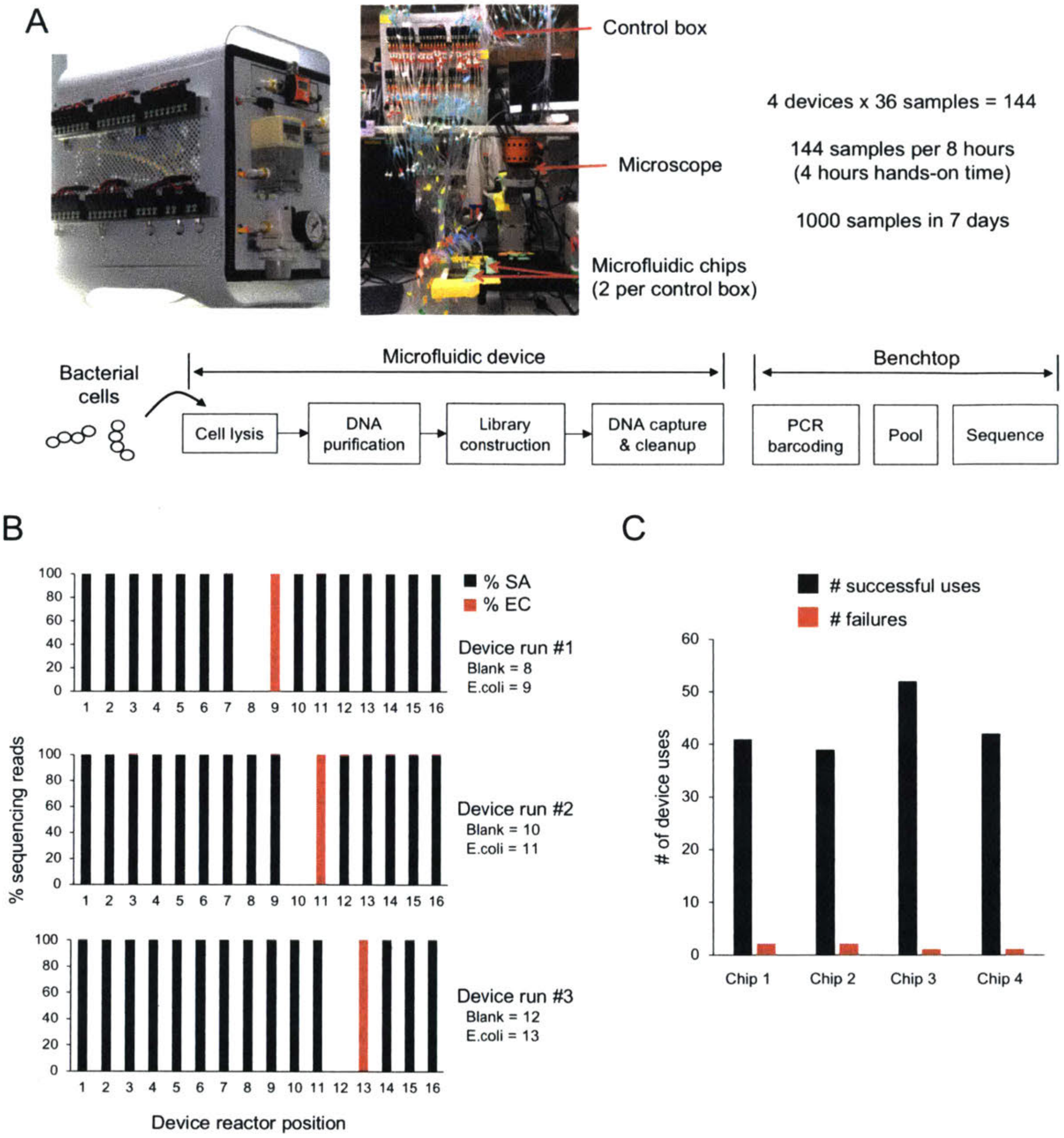
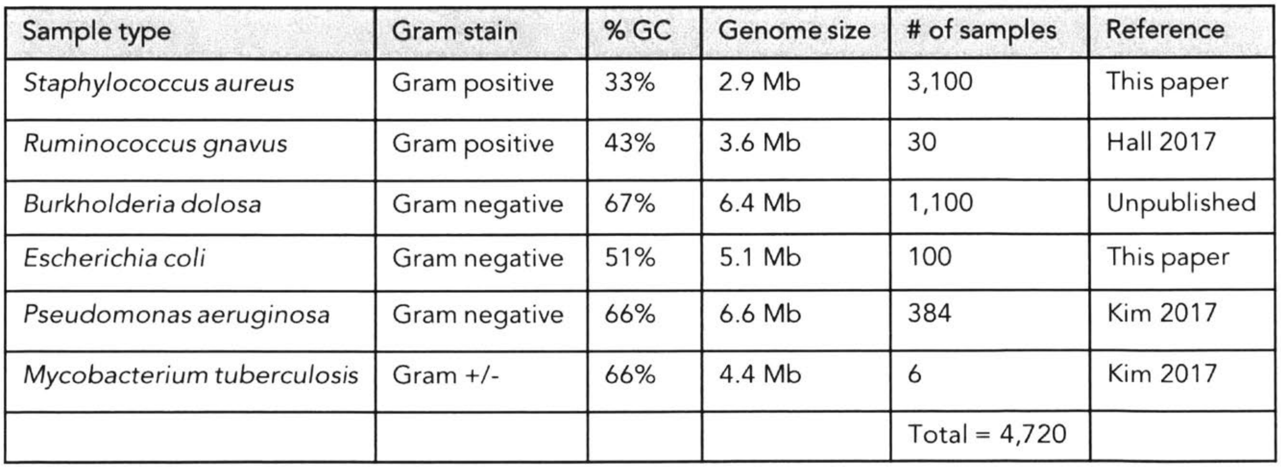


Figure 4-1. Microfluidic platform enables reliable high-throughput sequencing library preparation. A) Device setup includes control box (left) that connects to an air pressure source, computer, and the device. The computer opens and closes a series of valves on the control box that regulate air pressure to the microfluidic device. The microfluidic device uses this air pressure to perform all functions by pushing fluids or blocking their flow. Two devices are connected to each control box, and a microscope is used for viewing the device. B) Testing of cross-contamination between reactors was performed by including *S. aureus* (SA), *E. coli* (EC), and blank (no sample) side-by-side in the device. Cross-contamination was measured by the percent of sequencing reads that mapped to either SA or EC. Blank samples had nearly zero reads mapping to either genome. No sign of cross-contamination between run 1, run 2, and run3 (top, middle, and bottom, respectively) was detected. C) Microfluidic devices were reused reliably, with more than 40 back-to-back uses per device and minimal failed runs.

4.3.3 Input normalization and sample processing

We utilized our microfluidic device protocol to process over 4000 samples (Table 1). We were able to successfully process gram-positive and gram-negative bacteria, including *S. aureus*, *R. gnavus*, *B. dolosa*, and *E. coli*. All bacteria were lysed using the same lysis protocol, which includes a mixture of enzymes and chemicals. In preparing DNA libraries, each sample has to have the correct ratio of DNA to Nextera enzyme, as this determines the fragment size length. While other DNA library preparation protocols normalize DNA input by quantifying each sample and adjusting the concentration, this method would prove tedious and costly for thousands of samples. In Kim et al., DNA extraction and normalization for the 384 clinical isolates was performed outside of the device. In our new method, we normalized DNA input by standardizing the number of input cells per device reactor, taking advantage of the low per-reactor variation in cell lysis and DNA capture efficiency.



Sample type	Gram stain	% GC	Genome size	# of samples	Reference
Staphylococcus aureus	Gram positive	33%	2.9 Mb	3,100	This paper
Ruminococcus gnavus	Gram positive	43%	3.6 Mb	30	Hall 2017
Burkholderia dolosa	Gram negative	67%	6.4 Mb	1,100	Unpublished
Escherichia coli	Gram negative	51%	5.1 Mb	100	This paper
Pseudomonas aeruginosa	Gram negative	66%	6.6 Mb	384	Kim 2017
Mycobacterium tuberculosis	Gram +/-	66%	4.4 Mb	6	Kim 2017
				Total = 4,720	

Table 4-1: Thousands of bacteria successfully lysed and processed in device

4.3.4 Device sequencing libraries are highly reproducible

In order to utilize the device for high-throughput sample processing, data quality needs to be reproducible both for individual samples and across a large sample size. We wanted to demonstrate that (1) individual samples are reproducible in device processing to provide confidence for data accuracy, (2) genomic comparison between samples on the device and the bench are highly similar, and (3) data quality is highly reproducible across thousands of samples. Kim et al. already demonstrated that individual DNA libraries from the device are consistent across replicates and are no different from benchtop prepared samples with deep sequencing (see Figure 2C and 2D in Kim et al.). However, we wanted to validate this result given that we have a new protocol, are using *S. aureus* samples instead of *P. aeruginosa*, and are starting from whole cells instead of normalized genomic DNA as done by Kim et al.

First, to evaluate the reproducibility of individual sample processing, we prepared, sequenced, and analyzed three samples in triplicate. We aligned reads to a reference genome (USA300) and performed variant calling for each triplicate sample. Greater than 95% of single nucleotide polymorphisms (SNPs) were shared in the pairwise comparison of each triplicate sample (Figure 4-2A). Second, we validated the accuracy of the microfluidic device libraries by comparing whole genome sequences of 14 isolates prepared both on the device and with a standard benchtop method. For each duplicate sample, we compared *de novo* assemblies as well as SNP calling to a standard reference genome. *De novo* assembly metrics were highly similar with N50 difference <20kb at comparable sequence depths (Supp. Figure 4-1). Using reference-based mapping for variant calling, we again find that >95% of SNPs are shared between device and benchtop prepared samples (Figure 4-2B). All of the SNP differences from triplicate and benchtop comparison occurred in either low complexity regions or high SNP density regions of the genome which are challenging to resolve by the variant calling software.

Third, we tested data metrics across thousands of samples from the device and found that average coverage (180x), insert size (320bp), and duplication rate (0.2) are consistent across thousands of samples (Figure 4-2C). The distribution of insert size has a small peak near 200bp, attributable to a DNA cleaning step with beads that bind DNA only >200bp. Therefore, in the case of short fragment DNA libraries, there is a selection step that converts these to roughly 200bp. Overall, our sequencing metrics indicate that the starting input of whole cells from saturated culture can provide libraries with reproducible conversion efficiency and fragment size.

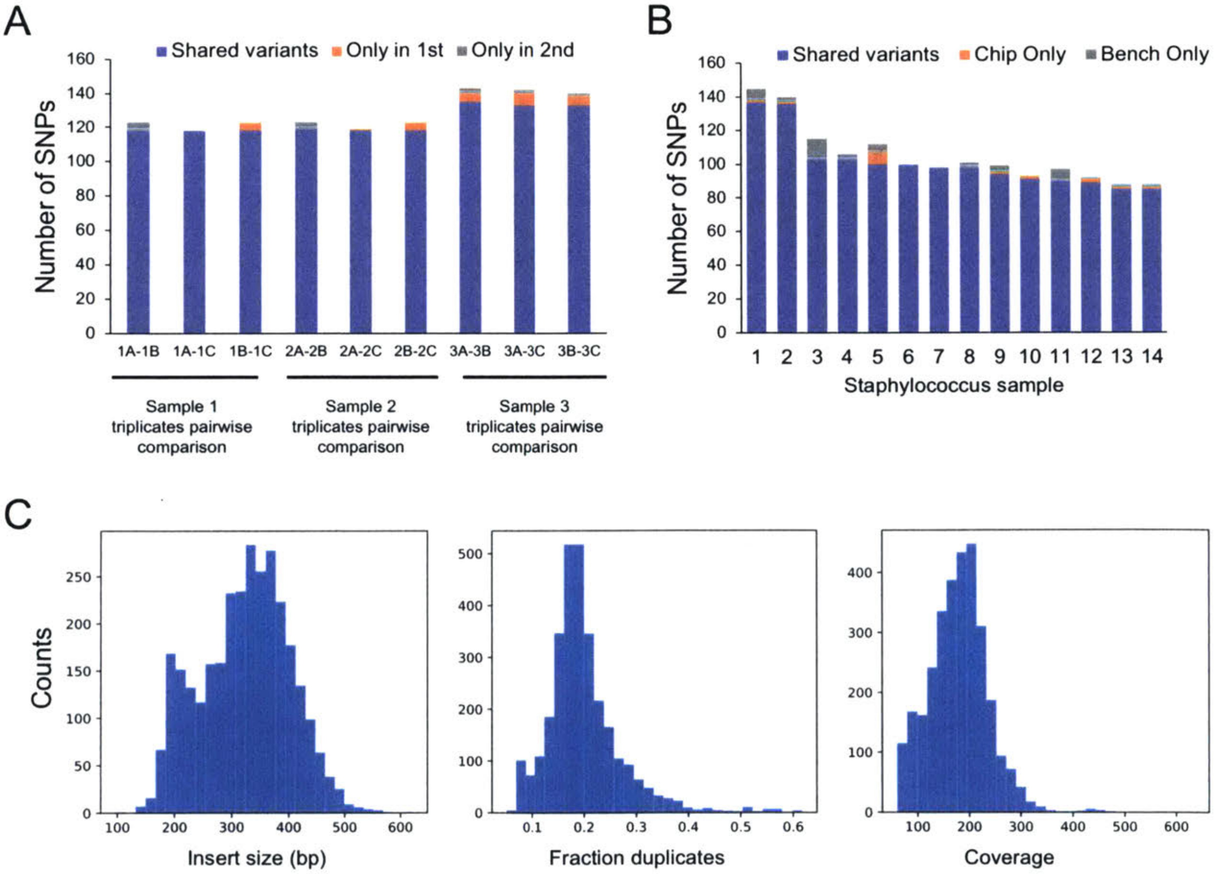


Figure 4-2: Device data quality is highly reproducible. A) Triplicate sample comparison on device shows >95% shared variants (SNPs). B) Same samples prepared on device and benchtop have >95% shared variants. C) Device sequencing metrics for insert size, duplicate fraction, and coverage are reproducible across 3000 samples.

4.3.5 Sequencing of 3000 MRSA isolates

Three thousand methicillin resistant *S. aureus* (MRSA) samples were collected in the Project CLEAR clinical trial to study MRSA transmission and recolonization rates (https://clinicaltrials.gov/ct2/show/NCT01209234). Samples were collected from hospitals across California from 820 patients over a period of 4 years. With our device, we sequenced all isolates and identified the sequence type with multilocus sequence typing (MLST), a standardized method that uses seven housekeeping genes to differentiate between strain types [159]. In our population, 45 different sequences types (ST) were identified; 34% were hospital-associated ST5 and 54% were community-associated ST8 (strain types common in the U.S.), with a smaller number of ST105 and ST30 (Figure 4-3A).

We also examined the staphylococcal cassette chromosome *mec* (SCC*mec*), a mobile genetic element that confers resistance to methicillin through the *mecA* gene. Using the *de novo* assemblies, we were able to identify six different SCC*mec* types, with 99% classified as SCC*mec* types II and IV (Table 4-2). Approximately 100 samples had contig breaks or other errors that prevented accurate typing. The majority of ST8 samples carry the SCC*mec* type IV, while ST5 strains had a mix of SCC*mec* type II and IV (Figure 4-3A). In addition, we found that about 200 samples contained additional SCC*mec* components on separate composite islands (without *mecA*). Previous studies have reported on non-*mecA* containing composite islands with ccrA and ccrB genes [160]–[162].

We screened the 3000 isolates for susceptibility to antibiotics using Mykrobe predictor, a genotype-based method to predict the resistance profile for 12 key antibiotics [163]. Consistent with the MRSA phenotype, the screening revealed all the isolates were genotypically resistant to methicillin and penicillin (Figure 4-3B), based on the presence of *mecA* and *blaZ* genes. For the rest of the panel, we observed more dynamic resistance genotypes. Erythromycin and ciprofloxacin were prevalent resistance genotypes at 89% and 88%, respectively, and clindamycin resistance was predicted in 56% of the isolates. A minority of the isolates are predicted to be resistant to gentamicin (11%) or mupirocin (10%), and less than 5% are resistant to rifampin, tetracycline, or trimethoprim. We found that predicted resistance to ciprofloxacin, clindamycin, and erythromycin correlated with SCC*mec* type II (Figure 4-3C). While samples with

SCC*mec* type IV had a mix of resistance and susceptibility to these three antibiotics (Table 4-3), SCC*mec* type II samples were nearly all resistant. The association of type II with resistance to these antibiotics has been found in other studies [164]. In particular, resistance to erythromycin is encoded by *ermA* in the element Tn*554*, which is carried on the type II SCC*mec* [165]. Overall, we found that samples had predicted resistance for an average of 4.5 antibiotics (range 1-9). The majority of samples (61%) have one of two antibiotic resistance profiles (antibiograms), but a diversity of resistance profiles exists for the remainder (Table 4-4). Upon examination of the fraction of STs per antibiogram, we find that antibiograms correlate with ST. This indicates that although individual antibiotic resistance does not correlate with ST (Table 4-4), patterns of resistance are strongly associated with ST. This association with ST in *Staphylococcus aureus* has also been reported previously [166].

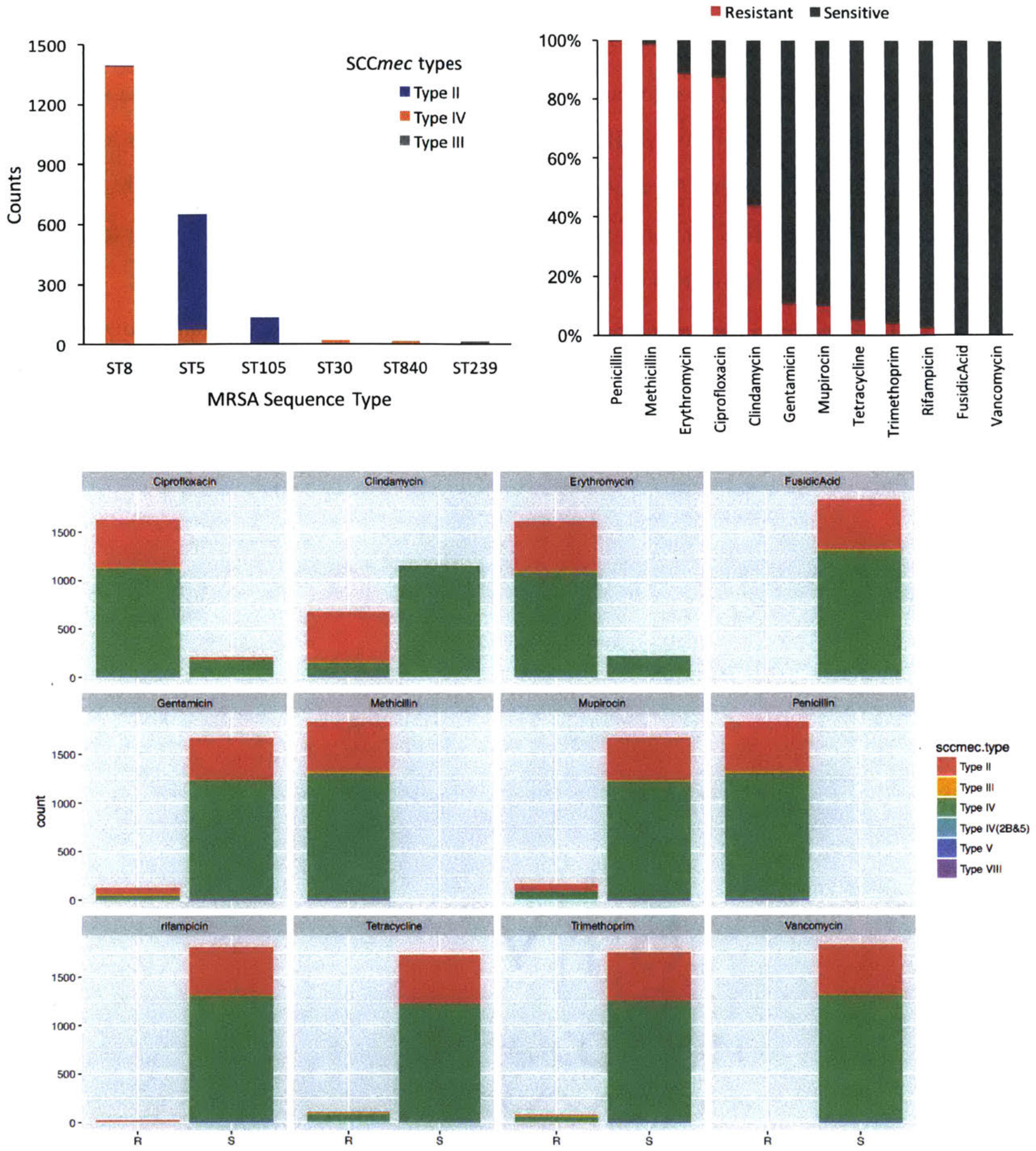


Figure 4-3: Sequencing of 3000 MRSA isolates. A) MLST was used to identify sequence types (ST), with ST5 and ST8 occurring most commonly. While up to 45 different sequence types were identified, only the top 6 are shown. Bars are colored by SCC*mec* type frequencies. B) Each MRSA sample was screened for antibiotic resistance to a panel of 12 drugs with Mykrobe, a genotype-based predictor of resistance. C) SCC*mec* type association with antibiotic resistance genotypes (R = resistant, S = sensitive).

Sample count	**SCC*mec* type**
1656	Type IV
757	Type II
15	Type III
12	Type V
2	Type IV(2B&5)
1	Type VIII

Table 4-2. SCC*mec* type frequencies across MRSA isolates.

Antibiotic resistance	**No. (%) of SCC*mec*-II strains (n=697)**	**No. (%) of SCC*mec*-IV strains (n=1512)**
Ciprofloxacin	659 (95%)	1303 (86%)
Clindamycin	695 (99.7%)	156 (10%)
Erythromycin	696 (99.9%)	1253 (83%)

Table 4-3. Antibiotic resistance genotypes in SCC*mec* type II and type IV samples.

Penicillin	Methicillin	Erythromycin	Ciprofloxacin	Clindamycin	Gentamicin	Mupirocin	Tetracycline	Trimethoprim	Rifampicin	Vancomycin	FusidicAcid	# predicted abx resistances	# of samples	Sequence Type - Counts					Sequence Type - Percent				
														ST5	ST8	ST105	ST30	Other	ST5	ST8	ST105	ST30	Other
R	R	R	R	R	R	S	R	R	S	S	S	8	14	1	0	0	0	13	7%	0%	0%	0%	93%
R	R	R	R	R	R	R	S	S	S	S	S	7	67	28	1	37	0	1	42%	1%	55%	0%	1%
R	R	R	R	R	R	S	S	S	S	S	S	6	94	75	0	8	0	11	80%	0%	9%	0%	12%
R	R	R	R	R	S	R	S	S	S	S	S	6	59	43	6	0	0	10	73%	10%	0%	0%	17%
R	R	R	R	R	S	S	S	S	R	S	S	6	45	42	0	1	0	2	93%	0%	2%	0%	4%
R	R	R	R	R	S	S	R	S	S	S	S	6	26	13	3	0	0	10	50%	12%	0%	0%	38%
R	R	R	R	S	R	R	S	S	S	S	S	6	17	0	17	0	0	0	0%	100%	0%	0%	0%
R	R	R	R	R	S	S	S	S	S	S	S	5	**736**	530	50	79	0	77	72%	7%	11%	0%	10%
R	R	R	R	S	S	R	S	S	S	S	S	5	47	0	47	0	0	0	0%	100%	0%	0%	0%
R	R	R	R	S	S	S	R	S	S	S	S	5	27	0	26	0	0	1	0%	96%	0%	0%	4%
R	R	R	S	R	R	S	S	S	S	S	S	5	24	20	0	1	0	3	83%	0%	4%	0%	13%
R	R	R	R	S	S	S	S	S	S	S	S	4	**1002**	2	931	0	0	69	0%	93%	0%	0%	7%
R	R	R	S	R	S	S	S	S	S	S	S	4	66	56	5	3	0	2	85%	8%	5%	0%	3%
R	R	S	R	S	S	R	S	S	S	S	S	4	15	0	15	0	0	0	0%	100%	0%	0%	0%
R	R	S	R	S	S	S	S	S	S	S	S	3	175	10	149	0	1	15	6%	85%	0%	1%	9%
R	R	R	S	S	S	S	S	S	S	S	S	3	120	3	110	0	1	6	3%	92%	0%	1%	5%
R	R	S	S	S	S	S	S	S	S	S	S	2	64	8	24	0	17	15	13%	38%	0%	27%	23%
R	S	S	S	S	S	S	S	S	S	S	S	1	16	2	0	0	1	3	13%	0%	0%	6%	81%

Table 4-4: Predicted antibiotic resistance profiles of MRSA samples (only for profile counts >14)

4.4 Discussion

Despite the increasing demand to sequence larger numbers of bacterial isolates, sequencing preparation methods have not kept pace with the growing need. Researchers and clinicians have large numbers of samples accessible and available to answer scientific questions or inform decisions in the clinic, yet there is no reliable, high-throughput, and low-cost.method for large-scale sequencing preparation. WGS has already been demonstrated as a useful epidemiological tool and has been used successfully to investigate hospital outbreaks. WGS has the capacity to characterize isolates with the highest resolution, predict resistance, and test for virulence. To help achieve the promise of large-scale WGS, we developed a DNA library preparation platform that can scale to thousands of bacteria, with a single researcher. Building upon the microfluidic design outlined in Kim et al., we have established a robust protocol that can process 1000 isolates in 7 days, starting from bacterial cells, in one integrated benchtop device.

Through parallel use of four 36-reactor devices, full protocol automation, device reuse with thorough cleaning, and high success rate of device runs, we prepared sequencing libraries directly from bacterial cells for 4000 samples (an order of magnitude higher than Kim et al.). We processed bacterial samples at a throughput of 144 samples per 7.5 hours, and we bypassed individual sample DNA normalization by loading each device reactor with an equal number of cells from saturated culture. The device protocol scales easily, and a wide range of batch sizes can be accommodated with minimal changes in cost or time, from 1 to 144 samples per day. Microbiology laboratories could process small batch sizes (e.g., 1-36 isolates) in a few hours and with <$50 in reagent costs. The maximum daily throughput achieved by a single researcher in this work was 288 samples per day (conducting two runs of four parallel devices), and with additional workers the only limitation on throughput is the number of microfluidic devices.

We utilized a sequencing-based assay to test for cross-contamination and showed that devices can be reused with high confidence. Overall, we were able to lyse and prepare DNA libraries from 6 different bacterial types, both gram-positive and gram-negative. The device sequencing data demonstrates high reproducibility across thousands of isolates and yields high quality sequence libraries that enable confident SNP detection. The device library preparation is robustly reproducible, as the data from this study and Kim et al. show high concordance of SNPs between replicate samples, even for different bacteria, different protocols, and different users. These data reliably facilitate strain typing, gene content analysis, SNP detection, and genotype-based prediction of antibiotic resistance profiles, demonstrating the utility for clinical microbiology labs and for infection control efforts.

Utilizing our device protocol, we processed 3000 MRSA isolates to demonstrate this method's applicability as a tool for molecular epidemiology and biology. We used the genomic data to identify sequence type, SCC*mec* type, and resistance to twelve antibiotics. We found that antibiotic resistance patterns correlate with sequence type, as ST5 and ST8 have distinctly different patterns of antibiotic resistance.

Our library preparation method provides an integrated workflow, incorporating all steps from cell lysis to DNA library construction. This procedure minimizes labor and use of consumables, as all steps occur within the microfluidic device. Other methods for scalable library preparation attempt to reduce reagent volume to drive down cost. One method for reduced-volume benchtop preparation scaled down the DNA library construction reaction to 2.5μL, starting with genomic DNA isolated from cells [154]. While cost is reduced due to lower reagent consumption, challenges still remain because manual pipetting is required and consumables (pipet tips, plates) are needed for each reaction. There is a floor for how low cost can go with pipetting-based reactions, as consumables remain a factor. Furthermore, low-volume pipetting can be inconsistent and cause reproducibility problems. Other methods use liquid handling machines to automate the library preparation and/or decrease reaction volume, and one example uses an acoustic liquid dispensing machine [157]. While automation and reduced volume minimize cost per reaction, these methods require the purchase of expensive equipment and consumables for each reaction. And in all cases, genomic DNA must be carefully normalized for each reaction. Microfluidics enable integration of all steps inside nanoliter-sized reaction chambers and use a system of valves and air pressure to move fluids around, eliminating the need for any consumables.

Our microfluidic technology is able to overcome many of the challenges with large-scale sample preparation, but there are still limitations. While the microfluidic device and controller system are made of inexpensive materials that can be scaled-up for low-cost production (order of magnitude less than a liquid handling machine), the system requires setup and basic training. The device currently does not have integration of DNA library barcoding amplification with large-scale processing, but the method has been successfully tested as small-scale and further development can incorporate this processing step at scale. Additionally, device design improvement can make the system more robust and reduce maintenance requirements, which will enable easier translation into other laboratories. A topic of ongoing work is to test the lysis protocol on a wider range of clinically relevant bacteria, as we tested only six bacterial types in this study. Furthermore, while this method provides for the "wet-lab" processing of samples, it is also important for clinicians and researchers to have access to streamlined data processing tools to enable rapid interpretation of WGS data.

Overall, the production of accurate WGS data from the device demonstrates the utility for clinical microbiology labs and infection control efforts. WGS enables genotype-based prediction of antibiotic resistance and high resolution isolate comparison with SNPs to infer transmission events. The microfluidic device protocol can be used to screen thousands of bacterial samples and applied in the clinic to monitor pathogen transmission, improve patient diagnosis, and track antibiotic resistance. This new ability to inexpensively and reliably process thousands of isolates also opens the door to bacterial genome-wide association studies, which aim to identify genetic variants that influence a specific phenotype. Our DNA library preparation protocol can put the power of high-throughput sequencing into the hands of clinicians and laboratory researchers.

4.5 Methods

4.5.1 Microfluidic device usage

Microfluidic devices were fabricated as described in Kim et al. While Kim et al. primarily used 16-reactor devices, we used 36-reactor devices in our study. To achieve a throughput of 144 samples per device run, four 36-reactor devices were used in parallel. A pair of devices was connected to a single controller box, for a total of two controller boxes that modulate air pressure to the microfluidic valves. Software written with Matlab was used to automate the device protocol, opening and closing valves to appropriately mix or dispense reagents as needed. A graphical user interface instructs the user to input reagents or start a thermal cycling program as needed, which was not present in Kim et al.

4.5.2 Making the device high-throughput; handling 3000 MRSA isolates

Five 36-reactor microfluidic devices were used to prepare all 3000 MRSA DNA libraries, as well as an additional 1000 other bacterial samples. No microfabrication was required besides creating the five devices during initial project fabrication. To prepare 3000 MRSA DNA libraries, approximately 100 device runs were required using the 36-reactor device. To maximize the throughput, 4 devices were used in parallel to achieve 144 samples per library preparation run. This enables processing of 3000 isolates in 3-4 weeks with a single researcher. Each library preparation run takes 7.5 hours, of which 4 hours are hands-on time. It is also possible to perform 2 device runs per day with 2 researchers working overlapping hours (15 hours from start to finish). This enables processing of 288 isolates/day.

The original library preparation protocol on the device in Kim et al. took approximately 10 hours, with 8 hours of hands-on time (for 16 samples total). In order to reduce protocol time, we automated the procedure, increased the throughput per run, reused devices, and made the protocol highly reliable and consistent. Hurdles to automation included potential for improper loading of enzymes, consistency across four devices running in parallel, unnoticed empty control lines, and device clogs. Steps were taken to address each challenge. In brief, we implemented user prompts through Matlab with clear instructions for each enzyme loading step. Microscopes were used to view and ensure proper enzyme loading. All solutions were filtered with a 5μm filter before loading into the device. Control checks were added at the start of device setup, to ensure that all four devices were fully functioning, and we used a very thorough clean-up protocol at the end to check for any final problems. Originally, fluid leakage from the tubing connected to the device led to valve malfunction; therefore, tubing connections were replaced with optimized metal pins and tubing that had a tighter fit and did not expand with heat. Finally, since many protocol steps require "deadfilling" a reactor on the PDMS device (pushing the air out through the pores of the PDMS and replacing that space with fluid), this often creates local air pressure in the device that can push air into control channels. To correct this, we added a low air pressure (5 PSI) to the control channels while in the "off" position (instead of 0 PSI), which enabled the control channels to resist filling with air during any deadfilling steps. The original success rate per chip run was approximately 75%; after automation and protocol changes this was increased to >95%.

4.5.3 Device reuse

Microfluidic devices were each reused an average of 40 times (1440 samples processed per device). To ensure that cross-contamination did not occur between device runs, we implemented a rigorous washing method. Two solutions were used: 10mM TrisHCl buffer with 1% Tween pH 8 and a KOH solution (0.04M KOH, 10mM DTT, 1% Tween in H_2O, pH 12.6). After a device run, all channels were filled with the KOH solution and incubated for 2 minutes. Channels were rinsed out with the TrisHCl buffer and dried with air. The KOH wash/TrisHCl wash and air dry was repeated once more. The final air dry continued until all channels were completely dry.

To test that cross-contamination did not occur between device runs, sequencing libraries with a negative control (no cells/DNA added), a positive control (*E. coli* cells), and 34 *S. aureus* samples were prepared on a single device. After every device run, the negative and positive controls were shifted to new reactors of the same device. This way, qPCR and sequencing of the negative and positive controls could be used to detect any cross-contamination. *E.coli* and blank sample sequencing reads were tested for presence of reads that mapped to staph.

4.5.4 Preparation of bacterial samples

MRSA isolates were prepared from frozen glycerol stocks. In order to normalize the cell input to the microfluidic device, each sample was grown to saturated culture at approximately 10^9 cells/mL. With an equal volume of sample added to the device, this would result in about 20,000 cells loaded into each 20 nL reactor. This way, each sample has approximately the same cell concentration and no downstream normalization is required. We recultured each isolate in 1 mL of LB for 16 hours at 37C on a plate shaker at 200 rpm. Samples were spun down at 4000 rcf for 5 minutes and resuspended in an equal volume of 5 μm-filtered PBS/25% glycerol mixture.

4.5.5 Sequencing library preparation on microfluidic device

To prepare DNA libraries, we loaded bacterial cells into the device, lysed cells, purified DNA with SPRI capture, fragmented and tagged DNA with sequencing adapters (Illumina Nextera protocol for tagmentation), added SDS to stop the tagmentation reaction, and eluted DNA libraries from the device. Kim et al. has full details of each protocol step. Changes that we made included the following: protocol changes to account for loading whole cells and performing lysis, lysis enzyme mixture was optimized for *Staphylococcus aureus*, incubation times were reduced to shorten the protocol, and the DNA purification step after tagmentation was omitted (see Supplemental Methods).

Briefly, bacterial cells were loaded into the device (4 μL per sample). A total of 20 nL of volume fills each reaction chamber. After all 36 samples were loaded, the device was placed on a thermal cycler and heated to 80C for 20 minutes (to "heat shock" the cells). Next, a mixture of enzymes and chemicals was added for lysis so that 20 nL fills the remaining half of the reaction chamber. The reaction was mixed for 10 minutes and then heated with a program of 37C for 45min, 50C for 30min, 75C for 5min, and 4C for 3min. Subsequently, a mixture of PEG, salt, and carboxylic acid beads was loaded into the device and mixed to precipitate the DNA in the high salt solution and cause DNA binding to the beads. Beads were captured on the device at a filter valve, creating a bead column that was then washed with 100% EtOH. DNA was eluted off of the beads and Nextera enzyme (Illumina) was added and mixed for 20 minutes to tagment the DNA. SDS was added to stop the reaction, and the final tagmented solution was eluted from the device in 10mM TrisHCl to a volume of 8 μL. 1 μL was used for qPCR testing (methods described in Kim et al.) and the remaining 7 μL were used for PCR barcoding and amplification.

4.5.6 Barcoding PCR and sample pooling

7 μL of tagmented product (~5-10pg) was mixed with 10 μL NEBNext High-Fidelity 2X PCR Master Mix (New England BioLabs), 1 μL 10μM forward primer, 1 μL 10μM reverse primer, and 1 μL PCR-clean water. Primers contained Illumina sequencing adapters and barcodes (used with permission from the Genomics Platform at the Broad Institute). We used a total of 40 forward and 16 reverse primers, yielding 640 unique barcode combinations (this was the maximum number of samples to pool in one sequencing lane). As HiSeq machines have separate lanes, samples with the same barcode can be loaded onto different lanes, allowing us to pool over 1200 samples onto a single HiSeq X. Samples were quantified using a SYBR green DNA dye assay. SYBR green dye (10,000X concentrate in DMSO, ThermoFisher Scientific) was diluted 1:10,000 with 10mM TrisHCl. 200 μL of this solution was mixed with 1 μL of PCR product, and the DNA amount was quantified with a fluorescence plate reader (using standards as a reference). Samples were pooled at equal volumes (regardless of DNA mass) and loaded onto a MiniSeq for initial sequencing to test how evenly the samples were pooled (mid-output 300 cycle kit). Reads were analyzed for presence of barcodes, and then samples were normalized and pooled based on relative abundance from MiniSeq reads to prepare a more even library pool for HiSeq. See Supp. Figure 4-2 for sequencing read distribution of samples directly from the device before normalization (MiniSeq barcode counts) compared to post-normalization (HiSeq barcode counts).

4.5.7 DNA sequencing

Samples were sequenced on the Illumina HiSeq2500 or HiSeqX, through the Broad Institute Genomics Platform. All runs were dual indexed 2x151bp or 2x101bp.

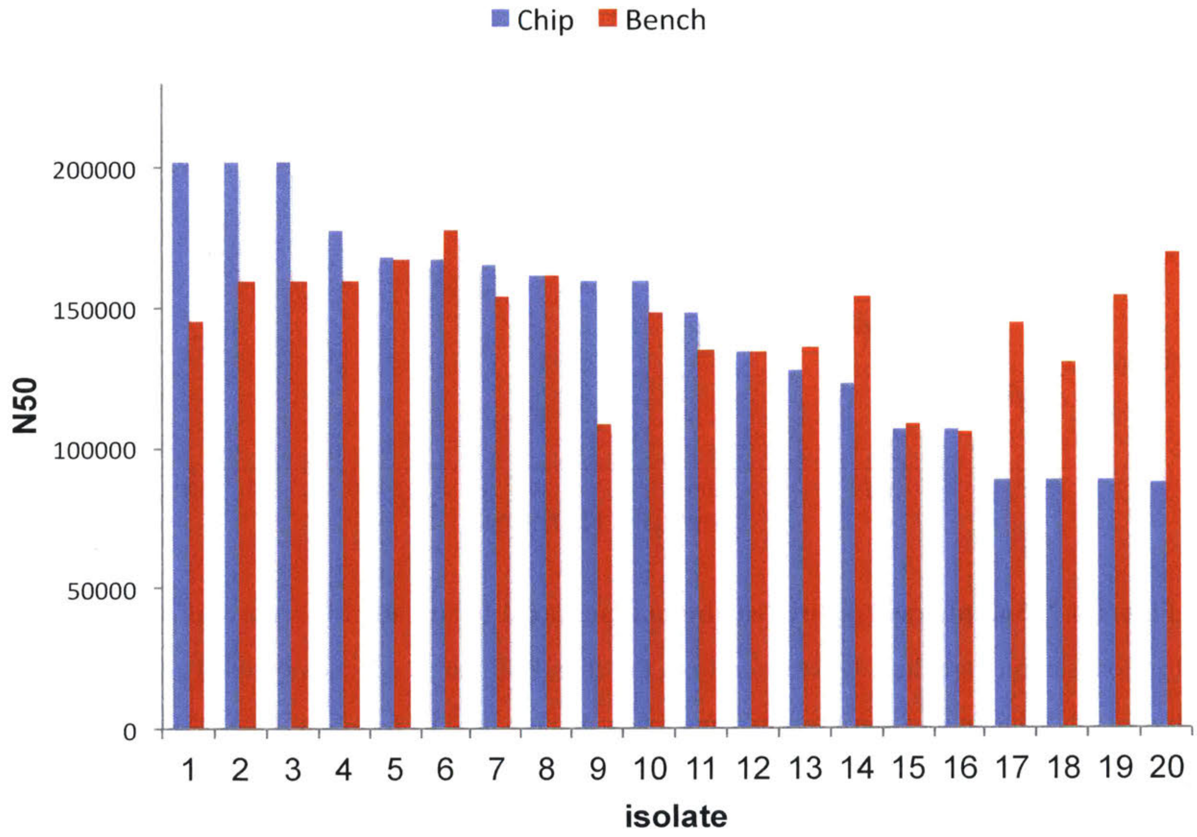
4.5.8 Genomic analysis

Genome assemblies were generated with SPAdes v3.9. Isolate sequence type (ST) and the corresponding clonal complex (CC) was determined using the PubMLST database (https://pubmlst.org/saureus/) and eBURST (http://saureus.mlst.net/eburst/). The SCC*mec* cassette typing was performed using pairwise blast to SCC*mec* elemets including the *ccr*, *mecA*, *mecI*, and *mecR* genes. Combination and orientation of SCC*mec* hits were parsed by custom scripts to infer SCC*mec* type (http://www.staphylococcus.net/). Sequencing and mapping metrics were calculated using Picard Tools (https://broadinstitute.github.io/picard/).

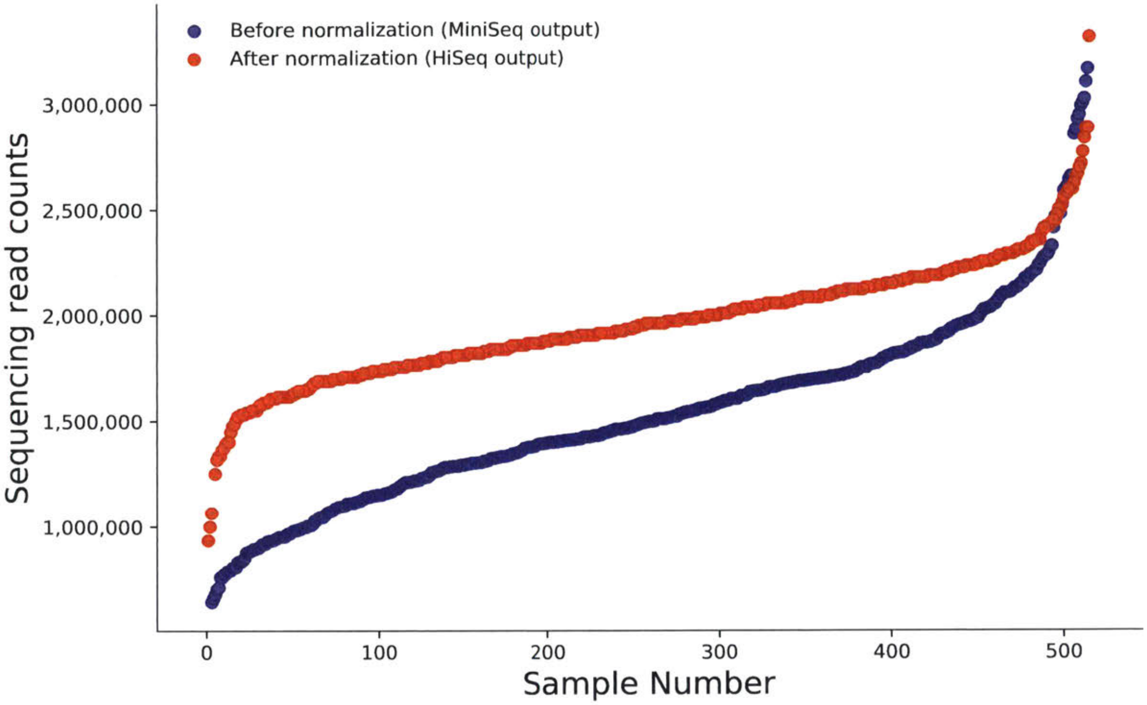
4.5.9 SNP calling and antibiotic resistance genotyping

Paired end reads were mapped using BWA v0.7.13 against a closed *S. aureus* chromosome of a matching sequence type. For the two most common ST8 and ST5 isolates, USA300-TCH1516 (NC_010079.1) and N315 (NC_002745.2), respectively, were used as reference genomes. SNP calling was performed using Pilon (https://www.broadinstitute.org/gaag/pilon). Duplicate reads were marked and ignored. Reads with a low mapping quality (<30), low coverage (<10 reads) or ambiguous variants from heterogeneous mappings were discarded. Antibiotic resistance genotype predictions were performed by Mykrobe software [163], using sequencing reads as input.

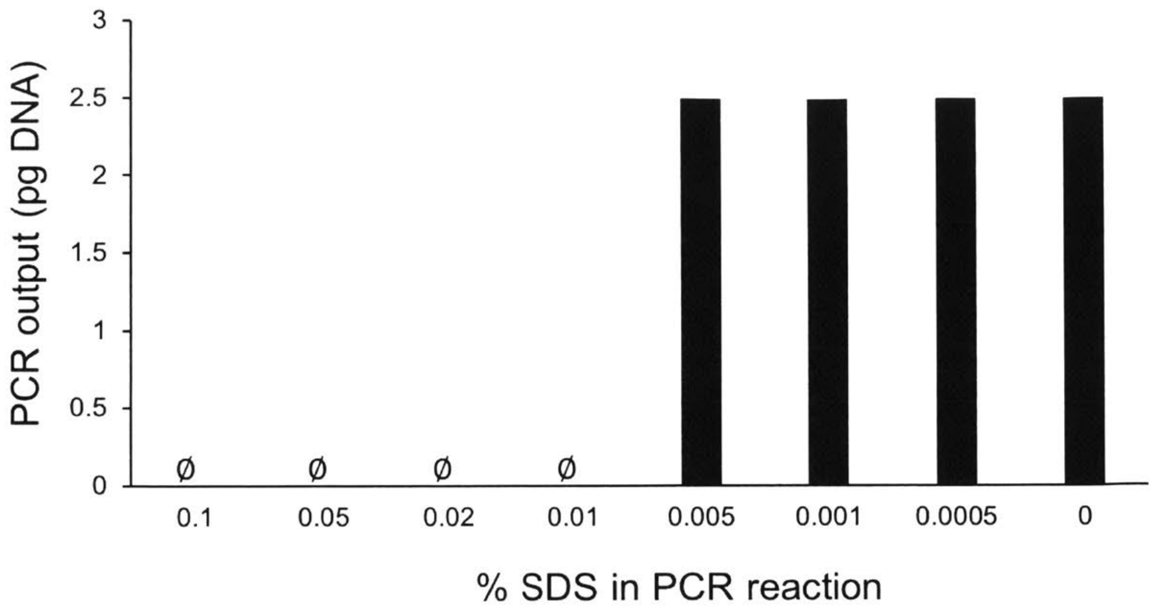
4.6 Supplemental Figures



Supp. Figure 4-1. ***De novo*** **assembly comparison between duplicate device and benchtop samples.** The N50 metric is compared between samples, which represents a weighted median of the assembled contigs. Given a set of contigs from the *de novo* assembly that are ordered from longest to shortest, the N50 length is defined as the contig length at which the sum of the contigs equal to and greater than this length is equal to 50% of the genome size. The same samples are prepared on the microfluidic device/chip (blue) and on the benchtop with standard methods (red).



Supp. Figure 4-2. Sample read count distribution measured before and after normalization. Before normalization, samples are directly removed from the chip, PCR barcoded, and pooled at equal volume regardless of DNA quantity. Sample sequencing reads are measured by MiniSeq (blue). After this initial sequencing, samples are normalized based on relative abundance and re-pooled for final HiSeq (red).



Supp. Figure 4-3. SDS inhibits PCR reaction with 0.01% or greater concentration. A titration experiment was performed to test the level at which SDS would inhibit a PCR reaction, using 2.5 pg of DNA for all reactions. PCR was performed as described in the methods, with a Nextera library and Nextera barcoding primers. SDS concentrations of 0.01% or greater resulted in no PCR output (as measured by DNA fluorescence and gel electrophoresis).

4.7 Supplemental Methods

4.7.1 Sequencing library preparation on microfluidic device (expanded)

First, bacterial cells were loaded into the device. 4 μL per sample was loaded into a pipette tip, the tip was plugged into one of the 36 input ports, and tubing with 10psi air flow was pushed into the top of the tip, forcing the sample into the device through dead-end filling. After all 36 samples were loaded, the device was placed on a thermal cycler and heated to 80C for 20 minutes ("heat shock"). Next, a mixture of enzymes and chemicals was added for lysis (SDS, proteinaseK, lysozyme, lysostaphin, mutanolysin, fragmentase, BSA). 20 nL of the lysis mixture was added to the 20 nL of cells and mixed for 10 minutes. Next the device was heated with a program of 37C for 45min, 50C for 30min, 75C for 5min, and 4C for 3min. Subsequently, 21 μL of a PEG solution (36% PEG, 2.5M KAc, and 0.5% Tween in water) was mixed with 4 μL of 2.8μm-diameter Dynabeads (M-270 Carboxylic Acid, Invitrogen) and 4 μL of 6um-diameter CML beads. This mixture was loaded into the device to fill 20 nL, pushing out half of the cell/enzyme mixture in the process. The reactors were mixed for 50 minutes to precipitate the DNA in the high salt solution and cause DNA binding to the beads. Beads were captured on the device at the filter valve, creating a bead column that was then washed with 100% EtOH. DNA was eluted off of the beads with a solution of 10mM TrisHCl and 0.1% Tween in water. Nextera enzyme (Illumina) was added and mixed for 20 minutes (6 μL Nextera was mixed with 1.5 μL of a 20x tagmentation solution), after which 4% SDS in water was added to stop the reaction. This final tagmented solution was eluted from the device in 10mM TrisHCl to a volume of 8 μL. 1 μL was used for qPCR testing (methods described in Kim et al) and the remaining 7 μL were used for PCR barcoding and amplification.

4.7.2 Cell lysis on microfluidic device

Prepare a lysis mixture of 10 μL P1A buffer (see below), 2 μL lysostaphin (Sigma, 2.5 mg/mL in 20mM NaOAc pH 5), 2 μL lysozyme (20 mg/mL in 10mM TrisHCl pH 8), 2 μL mutanolysin (Sigma, 6kU/mL in diH$_2$O), 2μL proteinase K (NEB, 20 mg/mL), and 2 μL SDS (10% in water). P1A buffer consists of

Qiagen P1 buffer, 10mM $CaCl_2$, 3mM $MgCl_2$, 0.5% Tween20, and 5% NP40. Combine 10 μL of the lysis mixture with 6 μL of dsDNA fagmentase (NEB) and 2 μL of BSA (NEB, 20 mg/mL, diluted 1:10 in P1A buffer). Mix and load 3.5 μL per device. Proceed with lysis protocol as outlined in Kim et al.

4.7.3 Removal of DNA purification step after tagmentation

Kim et al. performed a DNA purification step on the device after tagmentation and addition of SDS, in order to wash away SDS that may inhibit downstream processing. Normally, SDS inhibits a PCR reaction when present at a concentrations ≥0.01%. In our device reaction chambers, SDS is at 0.4% final concentration. However, upon elution of samples out of the device and into the 20 μL PCR reaction, the SDS concentration drops to 0.001%. Therefore, the SDS should not inhibit the subsequent PCR reaction. We performed SDS titration tests to verify that indeed PCR inhibition occurs at 0.01% and higher (Supp. Figure 4-3). Thus, our PCR reaction was not inhibited and we removed this DNA purification step from the protocol (saving time and reagents).

Chapter 5

Conclusion and future directions

This thesis describes the development and application of methods that enable sequencing of new types of microbial samples. As sequencing is no longer a bottleneck due to low costs and ease of use, sample preparation is now the limiting factor. Sequencing provides exciting new information about a system – what microbes are present, how they are genetically related to one another, and clues about how they function. With this information, it is possible to determine how and under what conditions microbes change, whether in a healthy commensal relationship or in a pathogenic context.

5.1 Investigation of the lung microbiome

In Chapter 2, I described methods for sensitive detection of microbial DNA in the healthy lung. The lung microbiome is a new field of study with many questions to explore, but the very low microbial biomass in the lung poses serious challenges. I developed an optimized protocol to selectively amplify microbial DNA while minimizing background noise, using mice as a model system. I found that bronchoalveolar lavage (BAL) of the lung provided the highest yield and most consistent sequencing result, and I applied a customized two-round PCR to amplify the 16S rRNA gene. A set of negative control samples was defined to diagnose common problems and identify potential contaminants. Through these methods, I profiled the lung microbiome of mice and found a number of common bacteria across samples. I identified that mice deficient in the adaptive immune system had a single bacterial taxon dominate their lung community, suggesting a role of the adaptive immune system in maintaining homeostasis of microbiota in the lung.

In Chapter 3, I applied these microbiome sampling methods to investigate the microbiome in the context of lung cancer. As previous research has shown that chronic lung disease can cause an alteration in the lung microbiome, I wanted to investigate how the microbiome might be altered in lung cancer and what implications this may have for disease progression. These are the first steps in teasing apart causation and correlation of microbiota and lung disease, as a main motivation in this field is determining whether the microbiota can be altered to promote a healthy state or prevent disease progression. In the work presented in Chapter 3, I showed that the lung microbiome has a reduced community diversity in tumor-bearing mice, and work with collaborators demonstrated that there was an increased bacterial load in tumor-bearing lungs. These findings suggest a relationship between the lung microbiome and lung tumor progression.

While investigation of the lung microbiome can have clear implications in health and the potential to contribute to methods for diagnosis, treatment, or prevention of lung disease, there are serious challenges and limitations in this area of research. First, collection of lung microbiome samples is unavoidably difficult – the lung must be accessed through invasive means either through the oral cavity and trachea or by surgical dissection of the thoracic cavity. While induced sputum is a common method for non-invasive sampling from the human lung, this method is not possible in mice and also introduced potential contaminants from the oral cavity. Lung lavage samples were identified as the preferred mouse lung collection method in this work. However, this method also proved difficult because lavage samples may be inconsistent in volume and the technique requires precision. It is also challenging to maintain perfect sterility when working with mice, as microbes cover the surface of their body. Furthermore, background bacterial DNA is hard to minimize; even with thorough optimization, background signal was detected from negative controls. Due to the low microbial biomass, many cycles of amplification are required which unavoidably amplify contaminants and introduce bias. While this thesis work describes methods to minimize these issues, the solutions are not perfect but instead provide incremental steps forward to support the field of lung microbiome research.

Importantly, the presence of bacterial DNA present in negative control samples is a difficult problem to address. While others in the field have often disregarded negative controls without presenting the sequencing results, we felt it was most appropriate to report and investigate all negative controls and carefully interpret the results. We found that many bacteria detected in negative controls were also common bacteria found in lung samples. And, importantly, these were also bacteria commonly found in the environment. The crux of the issue with lung microbiota detection is that bacteria found in the lungs are inherently common bacteria in our environment (like *Pseudomonas* and *Staphylococcus*) – the lung is populated by the bacteria we inhale and swallow and are regularly exposed to. Consequently, these bacteria are also ubiquitous in the lab environment and may appear as contamination. This is in contrast with gut microbiome research, where the majority of microbes are anaerobic and unlikely to be found living in the aerobic setting of the lab space and human skin. Researchers in this field need to recognize this overlap of bacteria and factor this into account when setting up controls for experiments and interpreting results. Of course, more accurate sequencing with longer read 16S rRNA or whole genome sequencing may genotypically differentiate similar bacterial taxa that are not distinguishable with short-length 16S sequencing, allowing for distinction between the species or strains present in the lung versus those in the lab environment.

The lung microbiome sequencing results in Chapter 2 and Chapter 3 indicate that even with many shared microbes between samples, few microbes appeared in a majority of samples and there was high variability in the relative abundance of microbes. This variability is a challenge when comparing samples and trying to derive a common signal from an experimental cohort. Furthermore, the microbiome is affected by caging, mouse vendor, gender, and other factors. Thus, it is important to have multiple experiments with mixed caging and to consider including perturbations (like antibiotics) to test for changes with respect to a control. This work was also limited by what was feasibly possible with mouse experiments – transfer of BAL between mice proved challenging and unsuccessful. Also, it would be ideal to culture bacteria from the lung to follow up with sequencing results and verify the viability of microbes. However, culture of the lung bacteria was attempted in many trials but yielded inconsistent and limited results. While the mouse provided a model system to test protocols and collect samples from a controlled environment, ideally these methods would be applied to human samples to move closer to true biological results in humans. Human lung sampling is challenging and can be invasive to patients, so it is helpful to optimize the processing protocols beforehand and maximize the accuracy of results.

5.2 Future directions of lung microbiome research

Future research in this field will benefit from continued optimization of the selective amplification of microbial DNA, establishment of common protocols in the field to allow researchers to compare results, perturbation experiments to alter the lung microbiota, and tracking of specific microbes. While the field of gut microbiome research has generally developed a consensus about extraction methods and 16S amplification procedures, the field of lung microbiome research has not. This is partly due to the fact that stool samples were included in the Human Microbiome Project and have been the focus of rigorous testing on sampling and processing methods. However, lung samples were not included in the HMP and the field is quite nascent. Therefore, further testing of protocols will help the field move towards commonly accepted practice in sample processing.

Specific to lung and tissue microbiome samples, the low microbial biomass still poses a challenge. Ongoing and future work will test methods for depleting mammalian DNA before selective amplification, as this may increase the signal to noise ratio and provide for a clearer, more robust signal. Since diverse research fields face this problem (lung microbiome, tissue microbiome, M. tuberculosis research, and viral sequencing from human samples, among others), ideally researchers can share protocols and build upon each other's work. In future lung microbiome studies, it will be critical to have a careful set of positive and negative controls, verifying both the technical reproducibility of the procedure as well as tracking sources of contamination. As described in Chapter 2, we believe future researchers should include a negative collection control, PCR control, positive sample control with a known bacterial mix, and multiple technical replicates. Making this a new standard would enable comparison of negative control data and procedural reproducibility across studies.

As the lung is a dynamic environment with a variable microbiome, future work may focus more on detecting changes from a steady state, such as introducing a perturbation like antibiotics. Ideally, monitoring the lung microbiome over time within one host would provide information about the dynamics of the microbiome. There is limited longitudinal research in healthy lung samples. However, further work in this area may help answer the simple yet tantalizingly evasive question: is there a stable, repopulating microbial community in the lung? How often do lung residents change and for how long do bacteria grow and reside in the lung? Answers to these questions will provide insight for how to interpret lung microbiome results and how to administer potential therapeutics.

Finally, new biological methods for tracking bacteria in the lung or depleting specific bacteria will provide major advancements to this research field. Bacteria may be introduced to the lung and tracked through fluorescence or DNA barcodes, enabling studies on colonization rates and bacterial growth properties in the lung. Additionally, methods for targeted depletion of bacteria may provide means to eliminate specific community members to examine their functional role, similar to the notion underlying gene knockout experiments in cells. Ultimately, this field may move towards highly specific tracking of microbial community members, allowing a more direct connection between microbes and host response.

The lung environment has been compared to "Life in Antarctica" – as in, the ecological landscape of the lung is similar to the harsh environment of Antarctica [167]. Just like humans are present in Antarctica but have no long-term colonization or growth, it is similarly hypothesized that microbes land in the lung but have limited growth capacity and residence time. However, when the lungs are in a diseased state, such as inflammation in the context of lung cancer or high mucus content in cystic fibrosis, the lung environment shifts and may have more amenable conditions due to changes in nutrient content, temperature, or immune surveillance. This landscape has much to be explored in the context of both health and disease.

5.3 Microfluidic sample preparation enables large-scale microbial sequencing

At the limits of scale, on the opposite end from the limits of sensitivity, preparing hundreds or thousands of bacterial samples for sequencing is a major hurdle. With the cost of sample preparation nearly 100-fold greater than sequencing itself, sample preparation is the bottleneck. Even with easy-to-detect bacteria and ample DNA, the sheer number of samples desired for large-scale clinical trials or molecular epidemiology studies surpasses existing capabilities for feasible sample preparation. In Chapter 4 of this thesis, I presented the development of a new platform for automating, miniaturizing, and streamlining the sample preparation process in a microfluidic device. Cells go into the device and prepared DNA libraries come out. With our technology, I was able to process thousands of bacterial samples, sequencing 3000 MRSA isolates from a clinical trial investigating decolonization methods. We screened for antibiotic resistance genes across the samples, and ongoing analysis will uncover the dynamics of patient colonization and recolonization with MRSA strains.

Whole genome sequencing enables the highest resolution for comparison of bacterial strains, down to a single nucleotide change. While this work aims to provide efficient WGS at large-scale, there are some limitations. Microfluidics are an excellent way to reduce reaction volumes and automate liquid handling steps that are normally required on the benchtop. However, this technology has challenges for adoption and ease-of-use. Complex microfluidic devices, like the one presented in this thesis, are difficult to manufacture and require fabrication facilities to initially develop and test. Even after a device design is finalized, microfluidic systems are "unconventional" in traditional science laboratories and initial training is needed for new users. Ideally, the device presented in this thesis could be packaged and produced at scale by a commercial entity for dissemination across many labs. However, there are a number of manufacturing challenges that have slowed the progress, and further development is needed to convert the microfluidic platform into a design that is feasibly produced at-scale. Furthermore, while we address the issue of generating WGS data at large-scale, another challenge is processing and interpreting large quantities of data. There are a number of bioinformatics challenges that must be tackled in order to rapidly interpret data and provide actionable results to researchers and clinicians.

5.4 Future directions of widespread microbial WGS

The drop in sequencing cost has powered a revolution in biology. Microbial DNA sequencing can track the spread of pathogens or allow us to investigate how microbes acquire antibiotic resistance. At large-scale, we can monitor pathogen outbreaks or screen thousands of patient samples retrospectively to examine the dynamics of pathogen transmission. To push the limits of scale in sequencing, microfluidics as well as liquid handling machines have miniaturized the process. Ongoing challenges are how to make these technologies more accessible for research laboratories and clinical laboratories. It is hard to predict how the sequencing landscape will change, as new technologies are constantly being developed that upend traditional methods. While Illumina next generation sequencing dominates, other methods for library preparation without the Illumina Nextera enzyme may provide avenues for sequencing preparation without this temperature sensitive and expensive enzyme. Additionally, new sequencing technologies such as nanopore sequencing may come into widespread use soon.

In the future, different applications may require different methods for sequencing. For example, in clinical oncology, high accuracy single nucleotide detection is needed to screen for human cancer mutations. On the other hand, testing for pathogens from a patient sample may likely only need a rapid yes/no answer for what pathogen is present. Therefore, speed and accuracy of sequencing will depend on the application. Furthermore, centralized sequencing facilities may be needed for certain applications, while portable, hand-held devices for rapid sequencing in the field may be needed for viral outbreak tracking.

For infectious disease testing in the clinic, whole genome sequencing may not necessarily be the future in pathogen detection. Although whole genome sequencing allows us to detect novel pathogens and perform very high resolution comparisons, clinicians often do not need the entire genome to take action. For example, polymerase chain reaction (PCR) assays that screen for a wide array of pathogens or specific subtypes are fast, easy to interpret, and low-cost. Genome sequencing has provided the knowledge to develop these PCR assays and identify DNA regions that are conserved and variable between samples, but in many cases identifying the genomic region of interest through PCR is sufficient. New methods for partial screening of the genome may also be developed with quick binary readouts. Likely, PCR assays and genome sequencing may serve complementary roles in the clinic.

Low-cost sequencing may turn the sequencer into the new “microscope” in biology. Just as genome editing has rapidly transformed the field of molecular biology with accessible and facile methods, sequencing is similarly turning into a ubiquitous tool in the lab. We can now read the DNA sequence of hundreds, or thousands, of microbes, at a few dollars per sample. Further advances in sample preparation and sequencing technology will bring us closer to fast, accurate, and scalable sequencing.

References

[1] N. C. Stenseth *et al.*, "Plague: Past, Present, and Future," *PLoS Med.*, vol. 5, no. 1, p. e3, Jan. 2008.

[2] A. L. Kozyrskyj, P. Ernst, and A. B. Becker, "Increased Risk of Childhood Asthma From Antibiotic Use in Early Life," *Chest*, vol. 131, no. 6, pp. 1753–1759, Jun. 2007.

[3] C. Braun-Fahrländer *et al.*, "Environmental Exposure to Endotoxin and Its Relation to Asthma in School-Age Children," *N. Engl. J. Med.*, vol. 347, no. 12, pp. 869–877, Sep. 2002.

[4] T. D. Lawley *et al.*, "Antibiotic Treatment of Clostridium difficile Carrier Mice Triggers a Supershedder State, Spore-Mediated Transmission, and Severe Disease in Immunocompromised Hosts," *Infect. Immun.*, vol. 77, no. 9, pp. 3661–3669, Sep. 2009.

[5] C. G. Buffie *et al.*, "Profound Alterations of Intestinal Microbiota following a Single Dose of Clindamycin Results in Sustained Susceptibility to Clostridium difficile-Induced Colitis," *Infect. Immun.*, vol. 80, no. 1, pp. 62–73, Jan. 2012.

[6] M.-C. Arrieta *et al.*, "Early infancy microbial and metabolic alterations affect risk of childhood asthma.," *Sci. Transl. Med.*, vol. 7, no. 307, p. 307ra152, Sep. 2015.

[7] S. Sengupta, M. K. Chattopadhyay, and H.-P. Grossart, "The multifaceted roles of antibiotics and antibiotic resistance in nature.," *Front. Microbiol.*, vol. 4, p. 47, 2013.

[8] R. Laxminarayan, "Antibiotic effectiveness: balancing conservation against innovation.," *Science*, vol. 345, no. 6202, pp. 1299–301, Sep. 2014.

[9] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–7, Dec. 1977.

[10] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 2, pp. 560–4, Feb. 1977.

[11] M. S. Rappé and S. J. Giovannoni, "The Uncultured Microbial Majority," *Annu. Rev. Microbiol.*, vol. 57, no. 1, pp. 369–394, Oct. 2003.

[12] C. Rinke *et al.*, "Insights into the phylogeny and coding potential of microbial dark matter," *Nature*, vol. 499, no. 7459, pp. 431–437, Jul. 2013.

[13] M. C. Wilson *et al.*, "An environmental bacterial taxon with a large and distinct metabolic repertoire," *Nature*, vol. 506, no. 7486, pp. 58–62, Feb. 2014.

[14] T. N. H. W. NIH HMP Working Group *et al.*, "The NIH Human Microbiome Project.," *Genome Res.*, vol. 19, no. 12, pp. 2317–23, Dec. 2009.

[15] R. Sender, S. Fuchs, and R. Milo, "Revised Estimates for the Number of Human and Bacteria Cells in the Body.," *PLoS Biol.*, vol. 14, no. 8, p. e1002533, Aug. 2016.

[16] C. Huttenhower *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, Jun. 2012.

[17] M. S. Donia *et al.*, "A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics," *Cell*, vol. 158, no. 6, pp. 1402–1414, Sep. 2014.

[18] M. Margulies *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, Sep. 2005.

[19] J. Shendure *et al.*, "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome," *Science (80-.).*, vol. 309, no. 5741, pp. 1728–1732, 2005.

[20] K. Wetterstrand, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)," 2017. .

[21] F. Syed, H. Grunenwald, and N. Caruccio, "Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition," *Nat. Methods | Appl. Notes, Publ. online 01 Novemb. 2009; | doi10.1038/nmeth.f.272*, Nov. 2009.

[22] Y. Feng, Y. Zhang, C. Ying, D. Wang, and C. Du, "Nanopore-based Fourth-generation DNA Sequencing Technology," *Genomics. Proteomics Bioinformatics*, vol. 13, no. 1, pp. 4–16, Feb. 2015.

[23] P. J. Brennan and H. Nikaido, "The Envelope of Mycobacteria," *Annu. Rev. Biochem.*, vol. 64, no. 1, pp. 29–63, Jun. 1995.

[24] P. E. Vandeventer *et al.*, "Mechanical disruption of lysis-resistant bacterial cells by use of a miniature, low-power, disposable device.," *J. Clin. Microbiol.*, vol. 49, no. 7, pp. 2533–9, Jul. 2011.

[25] R. McNerney *et al.*, "Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistance analysis: a call to action.," *Int. J. Infect. Dis.*, vol. 56, pp. 130–135, Mar. 2017.

[26] D. Nichols *et al.*, "Use of ichip for high-throughput in situ cultivation of &quot;uncultivable&quot; microbial species.," *Appl. Environ. Microbiol.*, vol. 76, no. 8, pp. 2445–50, Apr. 2010.

[27] C. L. Roose-Amsaleg, E. Garnier-Sillam, and M. Harry, "Extraction and purification of microbial DNA from soil and sediment samples," *Appl. Soil Ecol.*, vol. 18, no. 1, pp. 47–60, 2011.

[28] M. Castellarin *et al.*, "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.," *Genome Res.*, vol. 22, no. 2, pp. 299–306, Feb. 2012.

[29] L. T. Geller *et al.*, "Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine.," *Science*, vol. 357, no. 6356, pp. 1156–1160, Sep. 2017.

[30] J. R. Erb-Downward *et al.*, "Analysis of the lung microbiome in the 'healthy' smoker and in COPD.," *PLoS One*, vol. 6, no. 2, p. e16384, Jan. 2011.

[31] E. a Grice *et al.*, "Topographical and temporal diversity of the human skin microbiome.," *Science*, vol. 324, no. 5931, pp. 1190–2, May 2009.

[32] S. L. Salzberg *et al.*, "Next-generation sequencing in neuropathologic diagnosis of infections of

the nervous system.," *Neurol. Neuroimmunol. neuroinflammation*, vol. 3, no. 4, p. e251, Aug. 2016.

[33] R. Cotran, V. Kumar, T. Collins, and S. Robbins, *Robbins Pathologic Basis of Disease*. Philadelphia: Saunders, 1999.

[34] M. Hilty *et al.*, "Disordered microbial communities in asthmatic airways.," *PLoS One*, vol. 5, no. 1, p. e8578, Jan. 2010.

[35] E. S. Gollwitzer *et al.*, "Lung microbiota promotes tolerance to allergens in neonates via PD-L1," *Nat. Med.*, vol. advance on, May 2014.

[36] Y. Yun *et al.*, "Environmentally Determined Differences in the Murine Lung Microbiota and Their Relation to Alveolar Architecture," *PLoS One*, vol. 9, no. 12, p. e113466, Dec. 2014.

[37] M. A. Sze, J. C. Hogg, and D. D. Sin, "Bacterial Microbiome of lungs in COPD," *International journal of chronic obstructive pulmonary disease*, 2014. [Online]. Available: file:///Users/lagoudas/Downloads/COPD-38932-the-microbiome-in-copd_022114.pdf. [Accessed: 11-Nov-2014].

[38] P. C. Blainey, C. E. Milla, D. N. Cornfield, and S. R. Quake, "Quantitative Analysis of the Human Airway Microbial Ecology Reveals a Pervasive Signature for Cystic Fibrosis," *Sci. Transl. Med.*, vol. 4, no. 153, p. 153ra130-153ra130, 2012.

[39] M. J. Cox *et al.*, "Airway microbiota and pathogen abundance in age-stratified cystic fibrosis patients.," *PLoS One*, vol. 5, no. 6, p. e11044, Jan. 2010.

[40] J. Zhao *et al.*, "Decade-long bacterial community dynamics in cystic fibrosis airways.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 15, pp. 5809–14, Apr. 2012.

[41] C. U. Köser *et al.*, "Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak," *N. Engl. J. Med.*, vol. 366, no. 24, pp. 2267–2275, 2012.

[42] S. Reuter *et al.*, "Rapid Bacterial Whole-Genome Sequencing to Enhance Diagnostic and Public Health Microbiology," *JAMA Intern. Med.*, vol. 173, no. 15, p. 1397, Aug. 2013.

[43] A. Mellmann *et al.*, "Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting.," *J. Clin. Microbiol.*, vol. 54, no. 12, pp. 2874–2881, Dec. 2016.

[44] C. P. Harkins *et al.*, "Methicillin-resistant Staphylococcus aureus emerged long before the introduction of methicillin into clinical practice," *Genome Biol.*, vol. 18, no. 1, p. 130, Dec. 2017.

[45] J. A. Lees *et al.*, "Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration," *Elife*, vol. 6, p. e26255, Jul. 2017.

[46] S. G. Earle *et al.*, "Identifying lineage effects when controlling for population structure improves power in bacterial association studies," *Nat. Microbiol.*, vol. 1, no. 5, p. 16041, May 2016.

[47] M. Recker *et al.*, "Clonal differences in Staphylococcus aureus bacteraemia-associated mortality," *Nat. Microbiol.*, vol. 2, no. 10, pp. 1381–1388, 2017.

[48] R. P. Dickson, J. R. Erb-Downward, and G. B. Huffnagle, "Homeostasis and its Disruption in the

Lung Microbiome.," *Am. J. Physiol. Lung Cell. Mol. Physiol.*, p. ajplung.00279.2015, Oct. 2015.

[49] K. K. Barfod *et al.*, "The murine lung microbiome in relation to the intestinal and vaginal bacterial communities.," *BMC Microbiol.*, vol. 13, p. 303, 2013.

[50] K. K. Barfod *et al.*, "The Murine Lung Microbiome Changes During Lung Inflammation and Intranasal Vancomycin Treatment.," *Open Microbiol. J.*, vol. 9, pp. 167–79, 2015.

[51] M. Nguyen *et al.*, "The fermentation product 2,3-butanediol alters P. aeruginosa clearance, cytokine response and the lung microbiome," *ISME J.*, vol. 10, no. 12, pp. 2978–2983, Dec. 2016.

[52] Y. J. Huang and H. A. Boushey, "The microbiome in asthma," *J. Allergy Clin. Immunol.*, vol. 135, no. 1, pp. 25–30, 2015.

[53] T. Herbst *et al.*, "Dysregulation of allergic airway inflammation in the absence of microbial colonization.," *Am. J. Respir. Crit. Care Med.*, vol. 184, no. 2, pp. 198–205, Jul. 2011.

[54] T. Olszak *et al.*, "Microbial exposure during early life has persistent effects on natural killer T cell function.," *Science*, vol. 336, no. 6080, pp. 489–93, Apr. 2012.

[55] R. P. Dickson *et al.*, "The lung microbiota of healthy mice are highly variable, cluster by environment, and reflect variation in baseline lung innate immunity," *Am. J. Respir. Crit. Care Med.*, vol. 1487, no. 541, pp. 1–52, Mar. 2018.

[56] M. Kostric *et al.*, "Development of a Stable Lung Microbiome in Healthy Neonatal Mice," *Microb Ecol*, vol. 75, pp. 529–542, 2018.

[57] J. Scheiermann and D. M. Klinman, "Three distinct pneumotypes characterize the microbiome of the lung in BALB/cJ mice," *PLoS One*, vol. 12, no. 7, p. e0180561, Jul. 2017.

[58] N. Singh, A. Vats, A. Sharma, A. Arora, and A. Kumar, "The development of lower respiratory tract microbiome in mice," *Microbiome*, vol. 5, no. 1, p. 61, Dec. 2017.

[59] V. Poroyko *et al.*, "Alterations of lung microbiota in a mouse model of LPS-induced lung injury," *Am. J. Physiol. Cell. Mol. Physiol.*, vol. 309, no. 1, pp. L76–L83, Jul. 2015.

[60] S. Thoma-Uszynski *et al.*, "Induction of Direct Antimicrobial Activity Through Mammalian Toll-Like Receptors," *Science (80-.).*, vol. 291, no. 5508, pp. 1544–1547, Oct. 2001.

[61] A. O. Aliprantis *et al.*, "Cell activation and apoptosis by bacterial lipoproteins through toll-like receptor-2.," *Science*, vol. 285, no. 5428, pp. 736–9, Jul. 1999.

[62] A. Yoshimura, E. Lien, R. R. Ingalls, E. Tuomanen, R. Dziarski, and D. Golenbock, "Recognition of Gram-Positive Bacterial Cell Wall Components by the Innate Immune System Occurs Via Toll-Like Receptor 2," *J. Immunol.*, vol. 163, no. 1, pp. 1–5, Jun. 1999.

[63] W. J. Wiersinga *et al.*, "Toll-Like Receptor 2 Impairs Host Defense in Gram-Negative Sepsis Caused by Burkholderia pseudomallei (Melioidosis)," *PLoS Med.*, vol. 4, no. 7, p. e248, Jul. 2007.

[64] H. Zhang, J. B. Sparks, S. V Karyala, R. Settlage, and X. M. Luo, "Host adaptive immunity alters gut microbiota," *ISME J.*, vol. 9, no. 3, pp. 770–781, Mar. 2015.

[65] O. Takeuchi, K. Hoshino, and S. Akira, "Cutting Edge: TLR2-Deficient and MyD88-Deficient Mice Are Highly Susceptible to Staphylococcus aureus Infection," *J. Immunol.*, vol. 165, no. 10,

pp. 5392–5396, 2000.

[66] S. P. Preheim, A. R. Perrotta, A. M. Martin-Platero, A. Gupta, and E. J. Alm, "Distribution-based clustering: using ecology to refine the operational taxonomic unit.," *Appl. Environ. Microbiol.*, vol. 79, no. 21, pp. 6593–603, Nov. 2013.

[67] J. G. Caporaso *et al.*, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108 Suppl 1, no. Supplement 1, pp. 4516–22, Mar. 2011.

[68] S. J. Salter *et al.*, "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses," *BMC Biol.*, vol. 12, no. 1, p. 87, 2014.

[69] G. Biesbroek *et al.*, "Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection.," *PLoS One*, vol. 7, no. 3, p. e32942, Jan. 2012.

[70] M. Laurence, C. Hatzis, and D. E. Brash, "Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes.," *PLoS One*, vol. 9, no. 5, p. e97876, Jan. 2014.

[71] A. Remot *et al.*, "Bacteria isolated from lung modulate asthma susceptibility in mice," *ISME J.*, vol. 11, no. 5, pp. 1061–1074, May 2017.

[72] R. I. Adams, A. C. Bateman, H. M. Bik, and J. F. Meadow, "Microbiota of the indoor environment: a meta-analysis.," *Microbiome*, vol. 3, p. 49, Oct. 2015.

[73] P. Ganju *et al.*, "Microbial community profiling shows dysbiosis in the lesional skin of Vitiligo subjects," *Sci. Rep.*, vol. 6, no. 1, p. 18761, May 2016.

[74] A. Hoisington, J. P. Maestre, K. A. Kinney, and J. A. Siegel, "Characterizing the bacterial communities in retail stores in the United States," *Indoor Air*, vol. 26, no. 6, pp. 857–868, Dec. 2016.

[75] M. Camps-Bossacoma, F. J. Pérez-Cano, À. Franch, and M. Castell, "Gut Microbiota in a Rat Oral Sensitization Model: Effect of a Cocoa-Enriched Diet.," *Oxid. Med. Cell. Longev.*, vol. 2017, p. 7417505, 2017.

[76] L. O. Byerley *et al.*, "Changes in the gut microbial communities following addition of walnuts to the diet," *J. Nutr. Biochem.*, vol. 48, pp. 94–102, Oct. 2017.

[77] S. C. Di Rienzi *et al.*, "The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria.," *Elife*, vol. 2, p. e01102, Oct. 2013.

[78] R. P. Dickson *et al.*, "Bacterial Topography of the Healthy Human Lower Respiratory Tract.," *MBio*, vol. 8, no. 1, pp. e02287-16, Feb. 2017.

[79] G. R. Feehery *et al.*, "A method for selectively enriching microbial DNA from contaminating vertebrate host DNA.," *PLoS One*, vol. 8, no. 10, p. e76096, Jan. 2013.

[80] C. A. Marotz, J. G. Sanders, C. Zuniga, L. S. Zaramela, R. Knight, and K. Zengler, "Improving saliva shotgun metagenomics by chemical host DNA depletion.," *Microbiome*, vol. 6, no. 1, p. 42, Feb. 2018.

[81] N. Kumar *et al.*, "Efficient Enrichment of Bacterial mRNA from Host-Bacteria Total RNA Samples," *Sci. Rep.*, vol. 6, no. 1, p. 34850, Dec. 2016.

[82] M. Bisgaard, "Ecology and significance of Pasteurellaceae in animals.," *Zentralbl. Bakteriol.*, vol. 279, no. 1, pp. 7–26, Jun. 1993.

[83] J. G. Fox, *The Mouse in biomedical research. Volume II : Diseases*. Elsevier Academic Press, 2007.

[84] C. J. Czuprynski and A. K. Sample, "Interactions of Haemophilus-Actinobacillus-Pasteurella bacteria with phagocytic cells.," *Can. J. Vet. Res.*, vol. 54 Suppl, pp. S36-40, Apr. 1990.

[85] R. Dickson, J. Erb-Downward, C. Freeman, L. McCloskey, J. Beck, and J. L. Curtis, "Spatial Variation in the Healthy Human Lung Microbiome and the Adapted Island Model of Lung Biogeography," *Ann. Am. Thorac. Soc.*, vol. 12, no. 6, pp. 821–830, 2015.

[86] P. J. Turnbaugh *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, Jan. 2009.

[87] J. G. Caporaso *et al.*, "QIIME allows analysis of high-throughput community sequencing data," *Nat. Methods*, vol. 7, no. 5, pp. 335–336, 2010.

[88] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, p. 10, May 2011.

[89] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014.

[90] T. Z. DeSantis *et al.*, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.," *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–72, Jul. 2006.

[91] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010.

[92] Y. J. Huang *et al.*, "Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma," *J. Allergy Clin. Immunol.*, vol. 127, no. 2, p. 372–381.e3, 2011.

[93] M. A. Sze, J. C. Hogg, and D. D. Sin, "Bacterial microbiome of lungs in COPD.," *Int. J. Chron. Obstruct. Pulmon. Dis.*, vol. 9, pp. 229–38, Jan. 2014.

[94] A. A. Pragman *et al.*, "The lung tissue microbiota of mild and moderate chronic obstructive pulmonary disease," *Microbiome*, vol. 6, no. 1, p. 7, Dec. 2018.

[95] B. Coburn *et al.*, "Lung microbiota across age and disease stage in cystic fibrosis.," *Sci. Rep.*, vol. 5, p. 10241, Jan. 2015.

[96] S. M. Teo *et al.*, "The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development.," *Cell Host Microbe*, vol. 17, no. 5, pp. 704–15, May 2015.

[97] P. R. Marri, D. A. Stern, A. L. Wright, D. Billheimer, and F. D. Martinez, "Asthma-associated differences in microbial composition of induced sputum.," *J. Allergy Clin. Immunol.*, vol. 131, no. 2, pp. 346-52–3, Feb. 2013.

[98] M. a. Sze *et al.*, "The lung tissue microbiome in chronic obstructive pulmonary disease," *Am. J. Respir. Crit. Care Med.*, vol. 185, no. 10, pp. 1073–1080, 2012.

[99] M. Garcia-Nuñez *et al.*, "Severity-related changes of bronchial microbiome in chronic obstructive pulmonary disease.," *J. Clin. Microbiol.*, vol. 52, no. 12, pp. 4217–23, Dec. 2014.

[100] A. Jemal *et al.*, "Cancer Statistics, 2008," *CA. Cancer J. Clin.*, vol. 58, no. 2, pp. 71–96, Jan. 2008.

[101] A. Verdecchia *et al.*, "Recent cancer survival in Europe: a 2000–02 period analysis of EUROCARE-4 data," *Lancet Oncol.*, vol. 8, no. 9, pp. 784–796, Sep. 2007.

[102] M. J. Thun, S. J. Henley, D. Burns, A. Jemal, T. G. Shanks, and E. E. Calle, "Lung Cancer Death Rates in Lifelong Nonsmokers," *JNCI J. Natl. Cancer Inst.*, vol. 98, no. 10, pp. 691–699, May 2006.

[103] S. R. Pine *et al.*, "Increased Levels of Circulating Interleukin 6, Interleukin 8, C-Reactive Protein, and Risk of Lung Cancer," *JNCI J. Natl. Cancer Inst.*, vol. 103, no. 14, pp. 1112–1122, Jul. 2011.

[104] Y.-W. Cheng *et al.*, "Gender difference in human papillomarvirus infection for non-small cell lung cancer in Taiwan," *Lung Cancer*, vol. 46, no. 2, pp. 165–170, Nov. 2004.

[105] E. A. Engels, M. V. Brock, J. Chen, C. M. Hooker, M. Gillison, and R. D. Moore, "Elevated Incidence of Lung Cancer Among HIV-Infected Individuals," *J. Clin. Oncol.*, vol. 24, no. 9, pp. 1383–1388, Mar. 2006.

[106] A. J. Littman *et al.*, "Chlamydia pneumoniae infection and risk of lung cancer.," *Cancer Epidemiol. Biomarkers Prev.*, vol. 13, no. 10, pp. 1624–30, Oct. 2004.

[107] A. L. Laurila *et al.*, "Serological evidence of an association between Chlamydia pneumoniae infection and lung cancer.," *Int. J. cancer*, vol. 74, no. 1, pp. 31–4, Feb. 1997.

[108] H.-Y. Liang *et al.*, "Facts and fiction of the relationship between preexisting tuberculosis and lung cancer risk: A systematic review," *Int. J. Cancer*, vol. 125, no. 12, pp. 2936–2944, Dec. 2009.

[109] V. Søyseth, J. Š. Benth, and K. Stavem, "The association between hospitalisation for pneumonia and the diagnosis of lung cancer," *Lung Cancer*, vol. 57, no. 2, pp. 152–158, Aug. 2007.

[110] S. Kohno *et al.*, "The pattern of respiratory infection in patients with lung cancer.," *Tohoku J. Exp. Med.*, vol. 173, no. 4, pp. 405–11, Aug. 1994.

[111] Z. Y. Liu, X. Z. He, and R. S. Chapman, "Smoking and other risk factors for lung cancer in Xuanwei, China.," *Int. J. Epidemiol.*, vol. 20, no. 1, pp. 26–31, Mar. 1991.

[112] A. A. Santillan, C. A. Camargo, and G. A. Colditz, "A meta-analysis of asthma and risk of lung cancer (United States).," *Cancer Causes Control*, vol. 14, no. 4, pp. 327–34, May 2003.

[113] W. Zheng *et al.*, "Lung cancer and prior tuberculosis infection in Shanghai.," *Br. J. Cancer*, vol. 56, no. 4, pp. 501–4, Oct. 1987.

[114] M. W. Hinds, H. I. Cohen, and L. N. Kolonel, "Tuberculosis and lung cancer risk in nonsmoking women.," *Am. Rev. Respir. Dis.*, vol. 125, no. 6, pp. 776–8, Jun. 1982.

[115] A. Christopoulos, M. W. Saif, E. G. Sarris, and K. N. Syrigos, "Epidemiology of active tuberculosis in lung cancer patients: a systematic review," *Clin. Respir. J.*, vol. 8, no. 4, pp. 375–

381, Oct. 2014.

[116] G. D. Kirk *et al.*, "HIV Infection Is Associated with an Increased Risk for Lung Cancer, Independent of Smoking," *Clin. Infect. Dis.*, vol. 45, no. 1, pp. 103–110, Jul. 2007.

[117] M. J. Blaser and J. C. Atherton, "Helicobacter pylori persistence: biology and disease.," *J. Clin. Invest.*, vol. 113, no. 3, pp. 321–33, Feb. 2004.

[118] The EUROGAST Study Group, "An international association between Helicobacter pylori infection and gastric cancer," *Lancet*, vol. 341, no. 8857, pp. 1359–1363, May 1993.

[119] J. Parsonnet, G. D. Friedman, N. Orentreich, and H. Vogelman, "Risk for gastric cancer in people with CagA positive or CagA negative Helicobacter pylori infection.," *Gut*, vol. 40, no. 3, pp. 297–301, Mar. 1997.

[120] A. D. Kostic *et al.*, "Fusobacterium nucleatum Potentiates Intestinal Tumorigenesis and Modulates the Tumor-Immune Microenvironment," *Cell Host Microbe*, vol. 14, pp. 207–215, 2013.

[121] S. Bullman *et al.*, "Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer.," *Science*, vol. 358, no. 6369, pp. 1443–1448, Nov. 2017.

[122] N. Chaput *et al.*, "Baseline gut microbiota predicts clinical response and colitis in metastatic melanoma patients treated with ipilimumab," *Ann. Oncol.*, vol. 28, no. 6, pp. 1368–1379, Jun. 2017.

[123] A. E. Frankel *et al.*, "Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients.," *Neoplasia*, vol. 19, no. 10, pp. 848–855, Oct. 2017.

[124] V. Gopalakrishnan *et al.*, "Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients.," *Science*, p. eaan4236, Nov. 2017.

[125] V. Matson *et al.*, "The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients.," *Science*, vol. 359, no. 6371, pp. 104–108, Jan. 2018.

[126] B. Routy *et al.*, "Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors.," *Science*, p. eaan3706, Nov. 2017.

[127] S. H. Lee *et al.*, "Characterization of microbiome in bronchoalveolar lavage fluid of patients with lung cancer comparing with benign mass like lesions," *Lung Cancer*, vol. 102, pp. 89–95, Dec. 2016.

[128] H.-X. Liu *et al.*, "Difference of lower airway microbiome in bilateral protected specimen brush between lung cancer patients with unilateral lobar masses and control subjects," *Int. J. Cancer*, vol. 142, no. 4, pp. 769–778, Feb. 2018.

[129] G. Yu *et al.*, "Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features," *Genome Biol.*, vol. 17, no. 1, p. 163, Dec. 2016.

[130] J. E. Green and T. Hudson, "The promise of genetically engineered mice for cancer prevention studies," *Nat. Rev. Cancer*, vol. 5, no. 3, pp. 184–198, Mar. 2005.

[131] R. S. Herbst, J. V. Heymach, and S. M. Lippman, "Lung Cancer," *N. Engl. J. Med.*, vol. 359, no.

13, pp. 1367–1380, Sep. 2008.

[132] M. DuPage, A. L. Dooley, and T. Jacks, "Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase," *Nat. Protoc.*, vol. 4, no. 7, pp. 1064–1072, Jul. 2009.

[133] C. Urbaniak *et al.*, "Microbiota of human breast tissue.," *Appl. Environ. Microbiol.*, vol. 80, no. 10, pp. 3007–14, May 2014.

[134] C. Xuan *et al.*, "Microbial Dysbiosis Is Associated with Human Breast Cancer," *PLoS One*, vol. 9, no. 1, p. e83744, Jan. 2014.

[135] S. Pushalkar *et al.*, "The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression," *CANCER Discov.*, 2018.

[136] C. M. Lloyd and B. J. Marsland, "Lung Homeostasis: Influence of Age, Microbes, and the Immune System," *Immunity*, vol. 46, no. 4, pp. 549–561, Apr. 2017.

[137] L. N. Segal *et al.*, "Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype," *Nat. Microbiol.*, vol. 1, no. 5, p. 16031, May 2016.

[138] E. Bernasconi *et al.*, "Airway Microbiota Determines Innate Cell Inflammatory or Tissue Remodeling Profiles in Lung Transplantation," *Am. J. Respir. Crit. Care Med.*, vol. 194, no. 10, pp. 1252–1263, Nov. 2016.

[139] M. K. Shenoy *et al.*, "Immune Response and Mortality Risk Relate to Distinct Lung Microbiomes in Patients with HIV and Pneumonia," *Am. J. Respir. Crit. Care Med.*, vol. 195, no. 1, pp. 104–114, Jan. 2017.

[140] M. Gomes, A. L. Teixeira, A. Coelho, A. Araújo, and R. Medeiros, "The Role of Inflammation in Lung Cancer," Springer, Basel, 2014, pp. 1–23.

[141] R. P. Dickson, J. R. Erb-Downward, F. J. Martinez, and G. B. Huffnagle, "The Microbiome and the Respiratory Tract.," *Annu. Rev. Physiol.*, pp. 1–24, 2016.

[142] P. L. Molyneaux *et al.*, "The Role of Bacteria in the Pathogenesis and Progression of Idiopathic Pulmonary Fibrosis," *Am. J. Respir. Crit. Care Med.*, vol. 190, no. 8, pp. 906–913, Oct. 2014.

[143] R. P. Dickson, J. R. Erb-Downward, and G. B. Huffnagle, "Towards an ecology of the lung: new conceptual models of pulmonary microbiology and pneumonia pathogenesis.," *Lancet. Respir. Med.*, vol. 2, no. 3, pp. 238–46, Mar. 2014.

[144] A. Venkataraman *et al.*, "Application of a neutral community model to assess structuring of the human lung microbiome.," *MBio*, vol. 6, no. 1, p. e02284-14-, Jan. 2015.

[145] C. M. Bassis *et al.*, "Analysis of the Upper Respiratory Tract Microbiotas as the Source of the Lung and Gastric Microbiotas in Healthy Individuals," *MBio*, vol. 6, no. 2, pp. 1–10, 2015.

[146] R. P. Dickson, F. J. Martinez, and G. B. Huffnagle, "The role of the microbiome in exacerbations of chronic lung diseases," *Lancet*, vol. 384, no. 9944, pp. 691–702, Aug. 2014.

[147] J. M. Beck, V. B. Young, and G. B. Huffnagle, "The microbiome of the lung," *Transl. Res.*, vol. 160, no. 4, pp. 258–266, 2012.

[148] B. J. Marsland, A. Trompette, and E. S. Gollwitzer, "The gut-lung axis in respiratory disease,"

Ann. Am. Thorac. Soc., vol. 12, no. November, pp. S150–S156, 2015.

[149] D. N. O'Dwyer, R. P. Dickson, and B. B. Moore, "The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease," *J. Immunol.*, vol. 196, no. 12, pp. 4839–4847, Jun. 2016.

[150] X. Didelot, R. Bowden, D. J. Wilson, T. E. A. Peto, and D. W. Crook, "Transforming clinical microbiology with bacterial genome sequencing," *Nat. Rev. Genet.*, vol. 13, no. 9, pp. 601–612, Sep. 2012.

[151] World Health Organization, "ANTIMICROBIAL RESISTANCE Global Report on Surveillance," 2014.

[152] R. Dantes *et al.*, "National Burden of Invasive Methicillin-Resistant *Staphylococcus aureus* Infections, United States, 2011," *JAMA Intern. Med.*, vol. 173, no. 21, pp. 1970–1978, Sep. 2013.

[153] D. M. Aanensen *et al.*, "Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive Staphylococcus aureus in Europe.," *MBio*, vol. 7, no. 3, pp. e00444-16, May 2016.

[154] M. Baym, S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, and R. Kishony, "Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes," *PLoS One*, vol. 10, no. 5, p. e0128036, May 2015.

[155] S. Lamble *et al.*, "Improved workflows for high throughput library preparation using the transposome-based nextera system," *BMC Biotechnol.*, vol. 13, no. 1, p. 104, Nov. 2013.

[156] N. Rohland and D. Reich, "Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture.," *Genome Res.*, vol. 22, no. 5, pp. 939–46, May 2012.

[157] E. B. Shapland *et al.*, "Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process," *ACS Synth. Biol.*, vol. 4, no. 7, pp. 860–866, Jul. 2015.

[158] S. Kim *et al.*, "High-throughput automated microfluidic sample preparation for accurate microbial genomics," *Nat. Commun.*, vol. 8, p. 13919, Jan. 2017.

[159] M. C. Enright, N. P. Day, C. E. Davies, S. J. Peacock, and B. G. Spratt, "Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus.," *J. Clin. Microbiol.*, vol. 38, no. 3, pp. 1008–15, Mar. 2000.

[160] A. C. Shore *et al.*, "Characterization of a novel arginine catabolic mobile element (ACME) and staphylococcal chromosomal cassette mec composite island with significant homology to Staphylococcus epidermidis ACME type II in methicillin-resistant Staphylococcus aureus genotype ST22-MRSA-IV.," *Antimicrob. Agents Chemother.*, vol. 55, no. 5, pp. 1896–905, May 2011.

[161] A. C. Shore *et al.*, "DNA microarray profiling of a diverse collection of nosocomial methicillin-resistant staphylococcus aureus isolates assigns the majority to the correct sequence type and staphylococcal cassette chromosome mec (SCCmec) type and results in the subsequent id," *Antimicrob. Agents Chemother.*, vol. 56, no. 10, pp. 5340–55, Oct. 2012.

[162] K. Mongkolrattanothai, S. Boyle, T. V Murphy, and R. S. Daum, "Novel non-mecA-containing staphylococcal chromosomal cassette composite island containing pbp4 and tagF genes in a

commensal staphylococcal species: a possible reservoir for antibiotic resistance islands in Staphylococcus aureus.," *Antimicrob. Agents Chemother.*, vol. 48, no. 5, pp. 1823–36, May 2004.

[163] P. Bradley *et al.*, "Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis," *Nat. Commun.*, vol. 6, no. 1, p. 10063, Dec. 2015.

[164] A. Kilic, H. Li, C. W. Stratton, and Y.-W. Tang, "Antimicrobial susceptibility patterns and staphylococcal cassette chromosome mec types of, as well as Panton-Valentine leukocidin occurrence among, methicillin-resistant Staphylococcus aureus isolates from children and adults in middle Tennessee.," *J. Clin. Microbiol.*, vol. 44, no. 12, pp. 4436–40, Dec. 2006.

[165] N. Malachowa and F. R. Deleo, "Mobile genetic elements of Staphylococcus aureus," *Cell. Mol. Life Sci.*, vol. 67, no. 18, pp. 3057–3071, 2010.

[166] S. Kanjilal *et al.*, "Trends in antibiotic susceptibility in Staphylococcus aureus in Boston, Massachusetts, 2000-2014.," *J. Clin. Microbiol.*, p. JCM.01160-17, Nov. 2017.

[167] G. B. Huffnagle, R. P. Dickson, and N. W. Lukacs, "The respiratory tract microbiome and lung inflammation: a two-way street," *Mucosal Immunol.*, vol. 10, no. 2, pp. 299–306, Mar. 2017.