2017

Bacterial strain-tracking across the human skin landscape in health and disease

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

BACTERIA STRAIN-TRACKING ACROSS THE HUMAN SKIN LANDSCAPE IN HEALTH AND DISEASE

by

ALLYSON LINDSAY BYRD

B.S., University of Georgia, 2008

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader ____
W. Evan Johnson, Ph.D.
Associate Professor of Medicine and Biostatistics

Second Reader ____
Julia A. Segre, Ph.D.
Senior Investigator, NHGRI, NIH

Third Reader ____
Yasmine Belkaid, Ph.D.
Senior Investigator, NIAID, NIH

*"Step 1: Get your sh*t together.*

Step 2: Be amazing.

Step 3: Save the world."

-Belkaid lab motto

DEDICATION

I would like to dedicate this thesis to my parents, Ken and Kathy Byrd, whose journey over the last 4 years has been a constant inspiration.

ACKNOWLEDGMENTS

This thesis and my time in graduate school would likely not have happened without my first scientific mentor Daniel Promislow, who invited naïve, freshman me to work in his lab second semester of college. Although working with Drosophila was not my calling, he did ultimately direct me to the laboratory of Jessica Kissinger. Here under the guidance of the incredibly patient postdoc Jeremy DeBarry, I discovered my love for computational biology. Only after strong encouragement from Daniel, Jeremy, and Jessie, did I shoot for the stars and apply to graduate school outside of my comfort zone, the southeast.

After arriving in Boston, knowing no-one, I was surrounded by support from the entire BU bioinfo community. I would like to thank the 2012 starting class and Tank, for many fond memories in the cubes that first year. Special thanks to Joe Perez-Rogers for taking Challenge Project as seriously as I did. Thank you to Nacho Caballero for giving me an appreciation for the importance of an aesthetically pleasing presentation and teaching me the skills to make one. Thank you to my first year instructors who helped fill the gaps in my computational knowledge. Special thanks to Gary Benson and Evan Johnson for guiding Challenge Project and the creation of Clinical Pathoscope that ultimately served as the foundation of my graduate thesis. A warm thank you to Mary Ellen Fitzpatrick, Dave King, Johanna Vasquez, and Caroline Lyman for doing the behind the scenes magic that makes being in the bioinfo program extra special. Dave, I'm sorry for always having something wrong with my registration forms. Caroline and

Johanna, thank you for always inviting me to Boston and making me feel involved from 100s of miles away (plus the many nights at the Hotel Commonwealth were amazing!). Thank you to the rest of the BU students for the many memories of $1 beers at pub nights, team building and campfires at retreats, international conferences, and dance parties.

At the NIH, I would like to thank Phil Ryan and Phil Wang for working so hard to foster a graduate student community despite the NIH's many crazy rules and regulations. Next, I would like to thank everyone in the Segre lab who made all the metagenomic sequencing possible, Clay Deming and Cynthia Ng for developing the protocols, processing the hundreds of samples, and making extensive genome collections; Sean Conlan for making the sequences magically appear on the cluster and many coding sanity check conversations, Julia Oh for advice on analysis approaches, Shih-Queen Lee-Lin for extracting the highest quality RNA for the microarrays, and Heidi Kong for designing studies in such a clinically precise way the reviewers could never ding us. I'd also like to thank Dr. Kong and her team for recruiting, scheduling, and sampling the volunteers, without which none of these novel human datasets would have been possibly. Another thanks to Clay for his incredible isolate picking talents without which the immune potential of LM087 and aatyq may never have been discovered; thank you Sara Cassidy for noticing the puffiness of those mouse ears.

In building 4, I would like to thank the Belkaid lab for teaching me immunology and tolerating my naïve questions. A special thanks to Mike Askenase for helping me navigate the NIH and welcoming me into its grad student community, Nico Bouladoux for walking me through my first mouse experiments, Vanessa Ridaura for much microbiome (and occasionally shopping) related advice, Samira Tamoutounour for your beautiful innate immune panel and constant help setting up the machine, Seong-Ji Han for translating my thoughts into the most beautiful cartoons and organizing everyone to be at my thesis defense, Kim Beacht for removing the stress from takedown days with your speedy ear splitting, and finally Ollie Harrison for endless hours and many Saturdays in lab teaching me basic immunology skills, including scruffing mice, compensating/fixing the flow cytometer, analyzing flowjo, and practicing talks, but most importantly thank you for keeping me sane over these last few months. Thank you everyone who's shared your data with me so I could refine my computational skills outside of microbiome analysis and perfect the making of heatmaps. Thank you Belkaid lab for the numerous hilarious memories (many involving inappropriate quotes) of LPD Christmas videos, happy hours, dance parties, snatchy Christmas rats, and screaming birthday cakes.

Next, I would like to thank my committee members, Evan Johnson for advice that made the strain-tracking pipeline what it is today and signing numerous class registration forms, and Daniel Segre for making my committee meetings rather enjoyable conversations about science.

A huge thank you to my advisors Julie Segre and Yasmine Belkaid for allowing me to work with both of you and never keep a consistent schedule of my location. Although, it wasn't conventional, I feel so fortunate to have completed my PhD in 2 labs that are amazing at what they do! Julie, working in your lab with huge, novel human datasets was a computational biologists dream and allowed me to thrive in these last 3 years. Thank you for giving me numerous opportunities to present at conferences, review papers with you, and even write commentaries on a few; the discussions accompanying them were essential to my growth as a scientist. Yasmine, thank you for the opportunity to learn in your lab, because of it I am a more well rounded bioinformatist than most, and for allowing me to link my computational analyzes with functional experiments. Also, thank you for all the advice both in regards to science and life; I'll be seeking much more of it in the coming weeks.

Finally, I would like to thank my parents, siblings, and grandparents for all their support and patience over the last four years, and the many late night phone calls during walks home from lab.

BACTERIAL STRAIN-TRACKING ACROSS THE HUMAN SKIN LANDSCAPE IN HEALTH AND DISEASE

ALLYSON LINDSAY BYRD

Boston University Graduate School of Arts and Sciences

and

College of Engineering, 2017

Major Professor: W. Evan Johnson, Ph.D.

Associate Professor of Medicine and Biostatistics

ABSTRACT

Metagenomics, or genomic sequence of the community of microbiota (bacteria, fungi, virus), enables an investigation of the full complement of genetic material, including virulence, antibiotic resistance, and strain differentiating markers. The granularity to distinguish between closely related strains is important as within one species, these strains possess distinct functions and relationships to a host. To analyze metagenomic samples, I developed a reference-based approach that utilizes both single nucleotide variants and genetic content to assign species and strain-level designations. After refining this approach with complex simulated communities, I utilized it to analyze the microbial communities present in skin samples from healthy and diseased individuals.

First, to investigate strain-level heterogeneity in healthy adults, I focused on the common skin commensals *Propionibacterium acnes* and *Staphylococcus epidermidis* with well-documented sequence variation. Results indicated that an individual's strains of *P. acnes* are shared across multiple sites of his or her body, and that those strains are more similar within than between individuals. For *S. epidermidis*, in addition to individual site similarities, there were also site-specific strains. Overall these results emphasize that both individuality and site specificity shape our bodies' microbial communities. Based on longitudinal data, an individual's strain signatures remain stable for up to a year despite external, environmental perturbations.

I then used metagenomic data to explore microbial temporal dynamics in atopic dermatitis (AD; eczema), an inflammatory skin disease commonly associated with Staphylococcal species. Species-level investigation of AD flares demonstrated a microbial dichotomy in which *S. aureus* predominated on more severely affected patients while *S. epidermidis* predominated on less severely affected patients. Strain-level analysis determined that *S. aureus*-predominant patients were monocolonized with distinct *S. aureus* strains, while all patients had heterogeneous *S. epidermidis* strain communities. To assess the host immunologic effects of these species, I topically applied patient-derived strains to mice. AD strains of *S. aureus* were sufficient to elicit a skin immune response, characteristic of AD patients. This suggests a model whereby staphylococcal strains contribute to AD progression through activation of the host immune system. Overall, this strain-level analysis of healthy and disease communities provides previously unexplored resolution of human skin microbiome.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Ac	Antecubital crease
AD	Atopic dermatitis
Al	Alar crease
AMP	Antimicrobial Peptide
B	Baseline
Ba	Back
BU	Boston University
Ch	Cheek
DNA	Deoxyribonucleic Acid
Ea	External auditory canal
F	Flare
Gb	Glabella
Hp	Hypothenar palm
HV	Healthy Volunteer
Ic	Inguinal crease
Id	Interdigital web space
IFN	Interferon
IL	Interleukin
ISO	International Standards Organization
ITS	Intertranscribed spacer
Mb	Manubrium

N	Nares
NHGRI	National Human Genome Research Institute
NIAID	National
NIH	National Institutes of Health
NOD	Nucleotide Oligomerization Domain
Oc	Occiput
PAMP	Pathogen Associated Molecular Pattern
Pc	Popliteal crease
PF	Postflare
Ph	Plantar heel
PRR	Pattern Recognition Receptor
Ra	Retroauricular crease
RNA	Ribonucleic acid
rRNA	ribosomal RNA
SCORAD	Scoring Atopic Dermatitis
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
Tn	Toenail
TCR	T Cell Receptor
TLR	Toll-like Receptor
TMadv	Titi Monkey adenovirus
T_{reg}	T regulatory

TSB	Tryptic Soy Broth
Tw	Toeweb space
Vf	Volar forearm

CHAPTER 1 Introduction

1.1 Microbial studies in the past and present

Bacteria, fungi, and viruses are an integral part of ourselves, other host organisms and the environment in which we all live. Collectively referred to as the microbiome, these microbes remained our silent partners for most of history, until the 1760s when Dutch scientists Antoine Van Leeuwenhoek invented the first microscope and used it to observe the little "animalcules" living in the substance upon and between his teeth. Years later in 1885, Theodore Esherich studied diaper contents to understand why only a subset of his pediatric patients developed diarrhea in one of the original gut microbiome comparative studies (Escherich, 1988, 1989). Around this same time, German physician Robert Koch was formulating his infamous postulates as the gold standard criteria to establish a causative relationship between a microbe and a disease (Koch, 1890). Koch's original postulates can be summarized as follows: First, the microorganism occurs in every case of the disease; second, it is not found in healthy hosts; and third, after the microorganism has been isolated from a diseased organism and propagated in pure culture, the proposed pathogen can induce disease anew. In light of these postulates, for the years following, people focused on pathogenic microbes in regards to their roles in disease, while the benefits of commensal microbes were largely ignored. This was true in spite of the observation of 10x more bacterial than human cells in the body (Luckey, 1972; Savage, 1977), a ratio recently updated with more accurate numbers to 1:1 (Sender et al., 2016).

This imbalance was intrinsic to the available technologies of the time. Traditionally, microbial communities were explored with culture-based methods. Because this strategy favors microbes that thrive in artificial growth conditions, it underestimates the diversity of a community. In disease states, where a single pathogen often predominates, this approach is sufficient, almost preferable, to isolate the microbe and proceed with Koch's postulates. However, when studying our bodies' complex indigenous flora, culture-based approaches are less appropriate. Thus to capture the complete diversity, investigators began applying sequencing methods to characterize a community that circumvented the bottleneck of culturing. This approach utilizes amplification of the conserved small subunit ribosomal RNA genes (16S ribosomal DNA (rDNA)) as a taxonomic marker to identify bacterial members of microbial communities (Woese and Fox, 1977). Early on, this 16S rRNA method was used to show the diversity of flora in healthy individuals (Eckburg et al., 2005), the differences between obese versus lean twins (Ley et al., 2006), and describe the bacterial communities across body sites (Faveri et al., 2008; Gao et al., 2007; Hyman et al., 2005).

In 2001, the Human Genome Project was completed (Lander et al., 2001). Upon its successful completion, clinician scientists David Relman and Stanley Falkow advocated a continuation of the momentum with a "second genome project" to investigate our poorly understood indigenous microflora (Relman and Falkow, 2001). In 2007, following the success of several preliminary microbiome studies, a second genome project was realized with the NIH's funding of the Human Microbiome Project (HMP) initiative (Group et al., 2009). The main goals of the HMP were to utilize new high-throughput sequencing technologies to establish a baseline for normal volunteers, to make comparisons of microbes in health and disease, and finally provide data resources and analysis techniques to facilitate future microbiome studies. In total from this initiative, 4,788 samples from 242 'healthy' adults and five major body areas (oral, skin, nasal, gastrointestinal track as represented by stool, and urogenital) were used for 16S bacterial sequencing (Human Microbiome Project, 2012). From this large scale sequencing effort, they found that site was the strongest driver of community composition and that each site was predominated by characteristic phyla. However, within a site, there was a large amount of variation between individuals in the relative abundance of genera and species within those main phyla. With longitudinal sampling, it was shown that an individual's microbial composition is more similar to themselves over time than to others.

As the cost of sequencing a megabase of DNA has decreased from ~$5,000 to ~5 cents in the last fifteen years, the field of microbiome research has simultaneously bloomed from 103 papers in 2000 to 5,484 in 2014 [Figure 1.1], the majority of which have focused on gut microbes. In contrast to the heavily studied gut microbial communities, fewer studies have focused on the skin, a site where many disorders are similarly associated with an altered microbial state or dysbiosis. While several recent Reviews have focused on gut microbial ecology (Blaser, 2014; Donaldson et al., 2016), microbial communities of the skin in health and disease will be the focus of this introduction and the subsequent thesis.

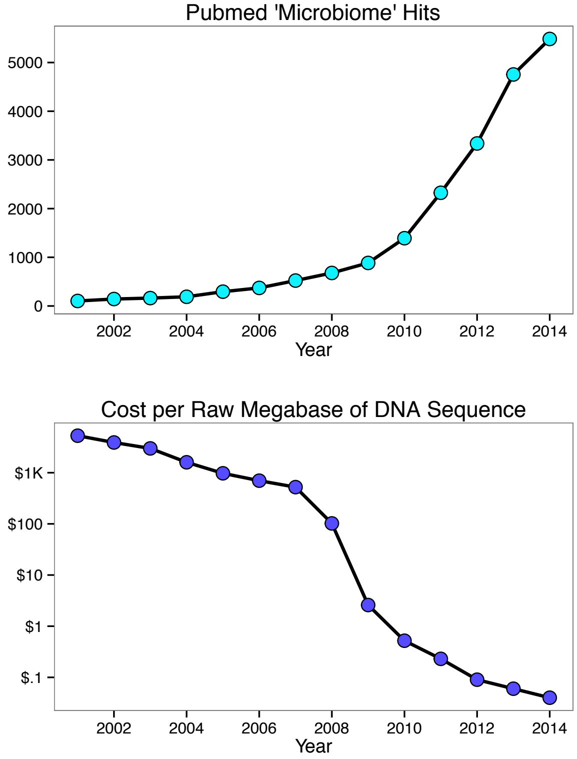


Figure 1.1. PubMed hits to "Microbiome" versus sequencing costs over time.

Top) Number of hits to "microbiome" in PubMed over time. Bottom) Cost to sequence a megabase of DNA over time (https://www.genome.gov/sequencingcostsdata/). The number of PubMed microbiome hits is positively associated with cheaper DNA sequencing**.**

1.2 Computational methods for microbiome analysis

A fundamental component of microbiome studies is classifying the organisms that are present. This was accomplished with culture-based methods until Woese and Fox found variations in ribosomal gene sequences could be utilized to classify microbes (Woese and Fox, 1977). In this original approach, ribosomal RNA from organisms was digested into individual oligonucleotides. The components of this oligonucleotide fingerprint were then Sanger sequenced and sequence variations within variable regions were used as molecular fingerprints to identify organisms. To identify and classify the constituents of bacterial communities the 16S ribosomal gene is used, while for fungal communities the internal transcribed spacer 1 (ITS1) region of the eukaryotic ribosomal gene complex is used to phylogenetically identify organisms (Schoch et al., 2012).

As sequencing technologies have advanced from Sanger sequencing to Roche/454 pyrosequencing and then Illumina, this original approach has continually adapted to accommodate increasing read depths and shorter read lengths. This has been accomplished with new primers for shorter amplicons, clustering methods to overcome sequencing error, and assembling methods to combine paired-end reads. With shorter amplicon lengths ~300 basepairs compared to >1,000, only a subset of the 16S gene can can be analyzed. This requires primers to be strategically placed within the 16S gene to optimize the diversity seen. Depending on the site being studied, different primer pairs may provide optimal results (Meisel et al., 2016). To date (Kuczynski et al., 2012), the primary pipelines for analyzing amplicon data are Mothur (Schloss et al., 2009) and Qiime (Caporaso et al., 2010). Both methods utilize a read clustering approach and subsequent comparison to curated reference databases to classify communities at a genera and when possible species level.

As the price of sequencing has fallen, the number of microbiome related studies has increased dramatically [Figure 1.1]. The majority of these studies were carried out with amplicon sequencing. However, over the last few years, the popularity of whole genome metagenomic sequencing has increased. With no targeted amplification, this method simultaneously captures all genetic material in a sample including human, bacterial, fungal, archeal, and viral, thus allowing relative kingdom abundances to be inferred

[Figure 1.2]. Because no marker gene is universally shared amongst viruses, viral community diversity is best captured in this way. Examining the complete microbial genome sequence provides sufficient granularity to differentiate strains within a species. The ability to differentiate strains is important as more and more studies highlight the functional differences that exist between strains within a species (Conlan et al., 2012; Tomida et al., 2013).

When designing experiments, it is important to consider that whole genome metagenomic sequencing is associated with additional costs (Franzosa et al., 2015). Financially, capturing the entire genomic content of a sample requires additional sequencing reads and is thus more expensive. This cost is particularly significant in the skin where depending on the body site sampled between 40 and 90% of reads are human (Oh et al., 2014). Thus samples must be sequenced deeply to reliably detect the microbial portion. In addition to sequencing costs, analysis costs should also be considered. Because few publicly available tools exist for metagenomic analysis, researchers need to independently create their own analysis approaches to explore the complexity of metagenomic data. The creations of such tools will be described throughout this thesis, but will be the primary focus of Chapter 2 and 3.

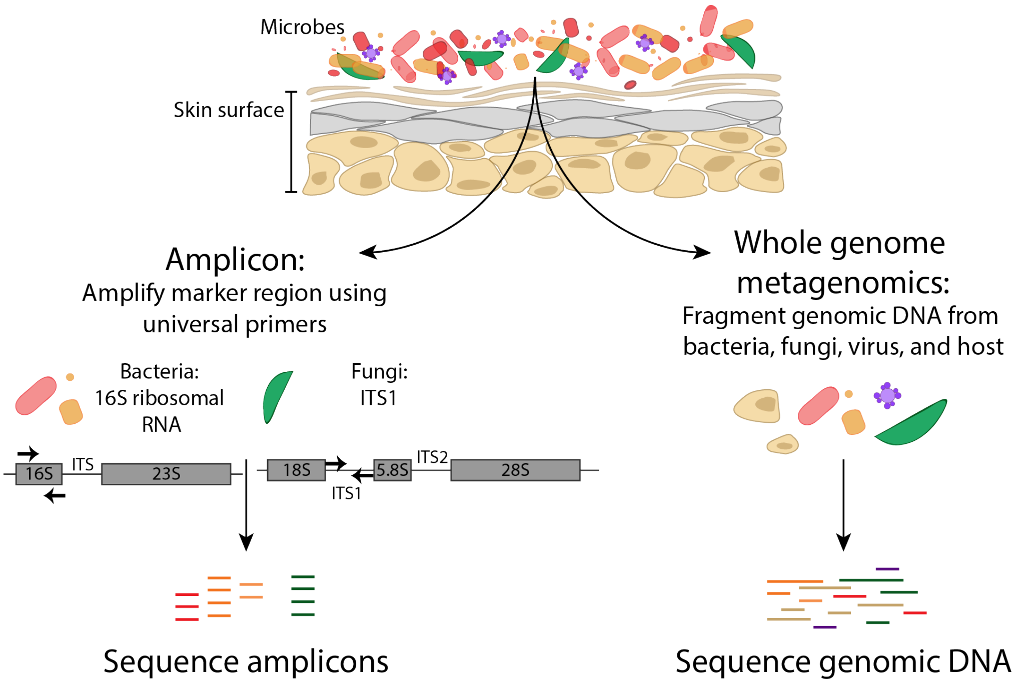


Figure 1.2. Amplicon versus whole genome metagenomic sequencing.

To study the microbial members of a community, two sequencing strategies can be utilized. Left) To amplicon sequence, primers are utilized to amplify conserved regions within a kingdom. For bacteria the 16S region of the ribosomal gene is utilized, while for fungi the ITS1 subunit is used. Right) Whole genome sequencing captures the entire complement of genetic material in a sample without a targeted amplification step.

1.3 Skin physiology shapes its microbial communities

Skin is the largest organ in our body composed of 1.8 meters squared of diverse habitats. It serves dual roles of acting as a physical barrier to outside pathogens while simultaneously providing a home to over 10^{10} resident bacterial cells (Belkaid and Segre, 2014). Structurally, the skin is composed of two distinct layers: the epidermis and dermis. The outermost layer, the epidermis, is composed of layers of ever-more differentiated keratinocytes. The top layer or stratum corneum is composed of terminally differentiated, enucleated keratinocytes, termed squames which are chemically cross-linked to each other to fortify the skin's barrier (Segre, 2006).

In addition to this conserved layer of structures, body sites provide diverse microenvironments varying in pH, temperature, moisture, sebum content, and topography (Grice and Segre, 2011). Based on these characteristics, sites can be grouped into these broad categories: sebaceous/oily(face, chest, and back), moist(bend of elbow, back of knee, and groin), and dry(volar forearm and palm). The environment of these sites is influenced by punctuation with varying densities of appendages such as sweat glands, hair follicules, and sebaceous glands. More abundant in moist sites, sweat glands are important for thermoregulation through the evaporation of water which also acidifies the skin making conditions unfavorable for the growth and colonization of certain microorganisms (Grice and Segre, 2011). In addition, sweat is laden with antimicrobial molecules, such as free fatty acids and AMPs that further inhibit microbial colonization (Gallo and Hooper, 2012). Connected to the hair follicle and more dense in oily sites, sebaceous glands secrete lipid-rich sebum, a hydrophobic coating that lubricates while also providing an antibacterial shield to hair and skin.

Despite the presence of lipid-rich sebum, skin is a nutrient desert compared to the nutrient rich environment of our intestines. Thus to survive in such a cool, acidic, desiccated environment, the resident microbes of our skin have adapted to utilize the resources that are available in sweat, sebum, and the stratum corneum (Scharschmidt and Fischbach, 2013). For example, facultative anaerobe, *P. acnes*, a prominent skin bacteria, is able to thrive in the anoxic sebaceous gland by using proteases to liberate the amino acid arginine from skin proteins (Holland et al., 1979) and lipases to degrade triglyceride lipids in sebum (Bruggemann et al., 2004). This degradation releases free fatty acids that promote bacterium adherence (Gribbon et al., 1993; Ingham et al., 1981; Marples et al., 1971). For mammals which produce smaller quantities of this triglyceride-rich sebum, *P. acnes* attaches less effectively and is thus found at lower abundances (Webster et al., 1981). The lipid-rich content of sebum and stratum corneum is also utilized by lipid auxotrophs Malassezia and Corynebacterium species, as they are unable to produce their own (Scharschmidt and Fischbach, 2013). Corynebacterium utilize these lipid compounds to generate corynemycolic acids that coat their cell surface (Scharschmidt and Fischbach, 2013). Consistent with skin's carbohydrate-deficient, lipid-rich environment, Malassezia genomes are enriched for lipases genes and depleted for carbohydrate utilizing ones compared to other fungi (Wu et al., 2015). Finally, Staphylococci have evolved many strategies for surviving on the skin, including the capability of being halotolerant, i.e. withstanding the high salt content of sweat, and utilizing the urea present in sweat as a source of nitrogen (Scharschmidt and Fischbach, 2013). To further promote colonization, various *Staphylococci* can also produce adhesions that promote attachment to the skin and proteases that are capable of liberating nutrients from the stratum corneum (Scharschmidt and Fischbach, 2013).

Given these specializations of bacteria for different niches it is not surprising that preliminary studies of the skin microbiome identified physiological characteristic as the strongest driver of bacterial community composition across body sites (Grice et al., 2009). Sebaceous sites are dominated by lipophilic Propionibacterium species, while humidophilic, halotolerant Staphylococcus and Corynebacterium species are abundant in moist areas, with higher densities of sweat glands [Figure 1.3]. In contrast, Malassezia is the dominant fungal genera across core body sites, while the feet harbor a more diverse community (Findley et al., 2013) [Figure 1.3]. These initial amplicon-based surveys will be expanded on with whole genome metagenomic samples in Chapter 3 and 4. In addition to utilizing valuable nutrients, our skin's bacterial communities are able to survive because of microbe/microbe interactions and a continuous dialogue with our immune system.

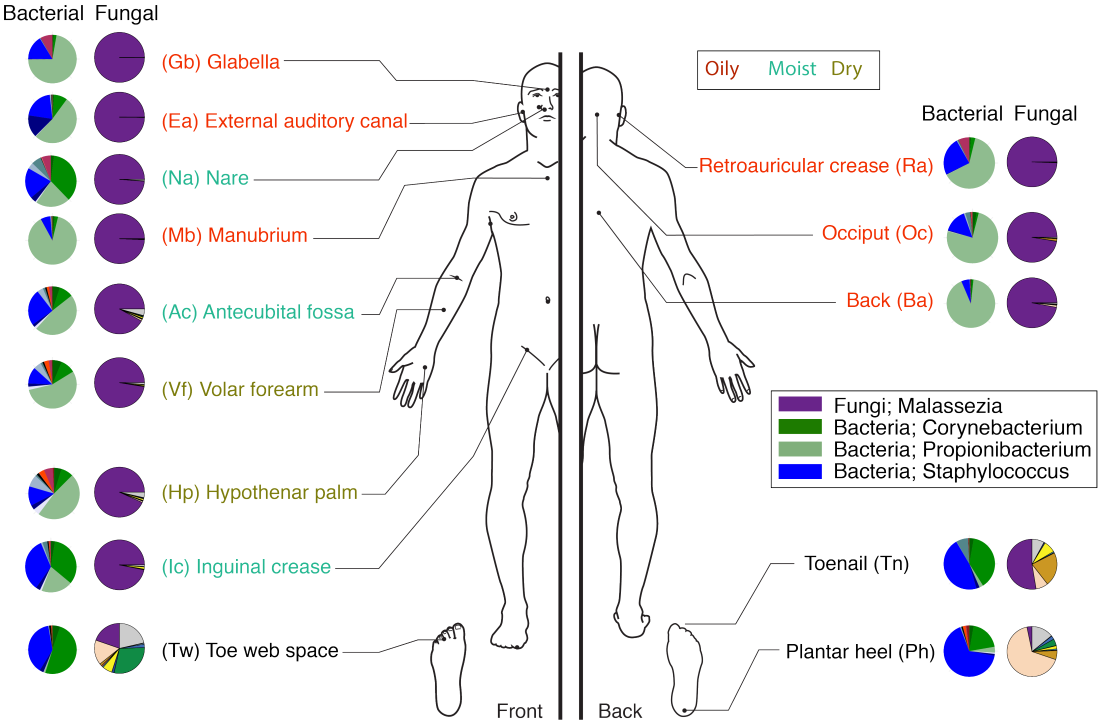


Figure 1.3. Physiological characteristics shape bacterial and fungal communities.

Consensus relative abundance plots show distribution of bacteria and fungi across different body sites based on previous 16S and ITS amplicon surveys. Site labels colored by microenvironment. Colors not shown in the microbial key may be grouped as 'Other'.

1.4 Interactions between cutaneous microbial species

In addition to host centric factors, community assembly and stability is driven by interactions between microbes. Microbes can act competitively to exclude one another or synergistically to optimize an effect. In the skin, *S. aureus* has been the focus of many colonization resistance studies. Colonizing the nares of 1/3 of the population, *S. aureus* presence is a significant risk factor for subsequent infections (von Eiff et al., 2001) (Weidenmaier et al., 2012). In clinical infections, 80% of *S. aureus* blood stream isolates match those identified in the patients nostril (von Eiff et al., 2001). Eradication of *S. aureus* in a surgical patient's nares strongly reduces their predisposition to invasive infection (Bode et al., 2010).

Because *S. aureus* frequently evolves resistance to antibiotics (DeLeo et al., 2010), alternate eradication strategies, particularly those that utilize indigenous microbes, are an active area of research (Pamer, 2016). These studies are akin to those exploring how soil microbes compete via antibiotic, bacteriocin production (Ling et al., 2015). Iwase and colleagues were the first to discover that a subset of *S. epidermidis* strains expressing the serine protease, Esp could inhibit *S. aureus* biofilm formation (Iwase et al., 2010). When Esp worked synergistically with the keratinocyte-produced AMP beta-defensin, *S. aureus* was completely inhibited [Figure 1.4]. Interestingly, 30 out of 30 sequenced *S. epidermidis* isolates encode the Esp gene (Conlan et al., 2012), but in Iwase's study only a subset were found to actually express it (Iwase et al., 2010). This discrepancy is an important reminder that coding potential does not guarantee expression. In a more recent study, Zipperer and colleagues found *S. lugdunensis* inhibited *S. aureus* growth via the productive of the antibiotic lugdunin, a novel thiazolidine-containing cyclic peptide (Zipperer et al., 2016)[Figure 1.4]. Importantly for long term therapeutic potential, after multiple generations *S. aureus* never developed resistance to either esp or lugdunin. This is in sharp contrast to traditional antibiotics which organisms rapidly evolve resistance to and emphasizes that naturally derived products will likely be a more effective means to block opportunistic pathogens. Notably, not all microbes inhibit *S. aureus;* Wollenburg et colleagues found that some Propionibacterium species could actually induce *S. aureus* aggregation and biofilm formation in a manner dependent on dose, growth phase, and pH (Wollenberg et al., 2014) [Figure 1.4].

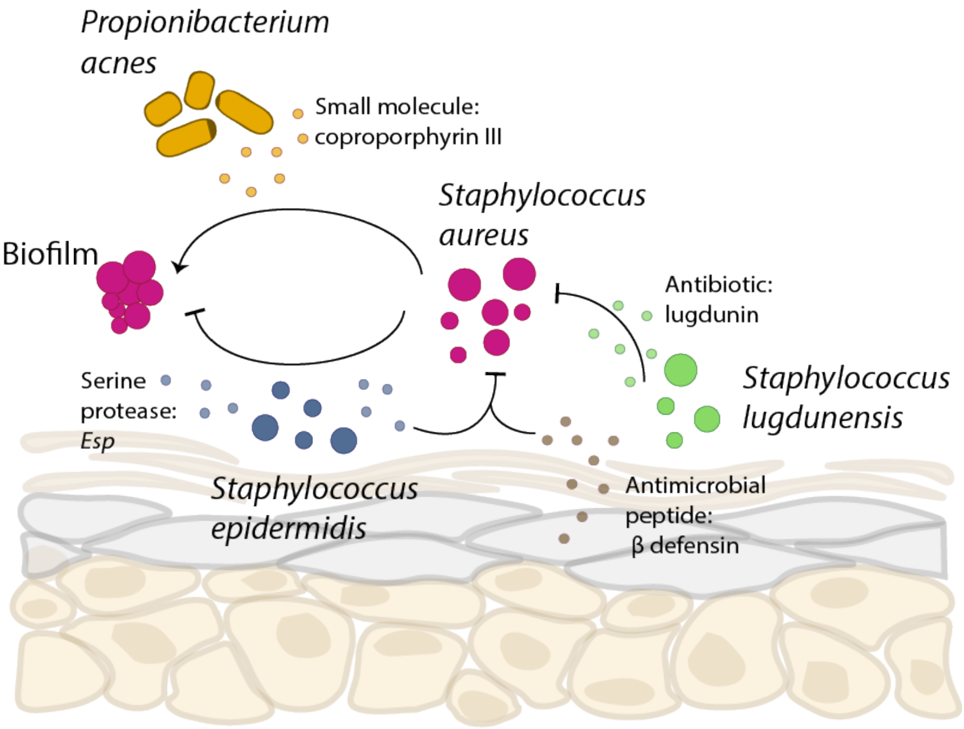

Propionibacterium acnes
Small molecule: coproporphyrin III
Staphylococcus aureus
Biofilm
Antibiotic: lugdunin
Staphylococcus lugdunensis
Serine protease: Esp
Staphylococcus epidermidis
Antimicrobial peptide: β defensin

Figure 1.4. Skin microbial communities are shaped by interactions between organisms.

In the skin, many interactions between commensals and *S. aureus* have been identified. The antibiotic, lugdunin, produced by *Staphylococcus lugdunensis*, prohibits colonization of *S. aureus*. Utilizing a different approach, *Staphylococcus epidermidis* can inhibit *S. aureus* biofilm formation with production of the serine protease, Esp. However, when Esp-expressing *S. epidermidis* acts in concert with keratinocytes expressing the antimicrobial peptide beta-defensin 2, *S. aureus* is effectively killed. In contrast to inhibiting *S. aureus*, *Propionibacterium acnes* produces a small molecule coproporphyrin III that promotes *S. aureus* aggregation and biofilm formation.

Other examples of competition between skin microorganisms also exist. Bomar and colleagues found *Corynebacterium accolens* could modify the local environment of the skin to inhibit growth of the contextual pathogen *Streptococcus pneumonia* (Bomar et al., 2016). This response was dependent on *C. accolens* using the lipase Lisp1 to release antibacterial free fatty acids from skin surface triacylglycerols. In another study, Christensen and colleagues performed pairwise antagonism assays with isolates from their culture collections of *S. epidermidis* and *P. acnes* isolates (Christensen et al., 2016). One clade of *P. acnes* exhibited a higher antimicrobial activity against *S. epidermidis*, likely due to a thiopeptide conserved between genomes in the clade. In the reverse, the majority of tested *S. epidermidis* strains were capable of inhibiting *P. acnes* in vitro. Computationally, they predicted a variety of different elements that could be responsible in different strains. In an even broader study of 89 Staphylococcus isolates from 6 species it was found that 84% could produce antimicrobial substances against common skin bacteria, indicating that bacteriocin capacity is an essential trait among skin commensals (Janek et al., 2016).

In contrast to many species/species interaction studies, investigations of dynamics between strains within a species are more rare. Extrapolations from metagenomic data have revealed two patterns of strain colonization within a species. For some species, there is a single dominant strain, while for other species, multiple strains coexist. In the gut of infants, the species *Escherichia coli, Faecalibacterium prausnitzii, Bacteroides fragilis*, and *Haemophilus parainfluenzae* exist as a single predominant strain while *Bacteroides vulgatus* exists as a heterogeneous community (Yassour et al., 2016). In an animal model of strain competition, it was shown that germ free mice mono-associated with a single *B. fragilis* isolate were resistant to colonization by a different *B. fragilis* strain but susceptible to colonization by a separate Bacteroides species (Lee et al., 2013). Interestingly, competition exclusion was also demonstrated by 3 other Bacteroides species as the initial colonizer, but *E. coli* did not exhibit this characteristic. While the heterogeneous *B. fragilis* strain communities observed in the babies' guts conflict with the competition exclusion observed in animal models, this data emphasizes that although reductionist germ-free experiments are a great place to start, they may not always reflect the ecological dynamics of a more complex community.

As will be described more in Chapter 4, in the skin *P. acnes* and *S. epidermidis* exist as stable heterogeneous communities of strains (Oh et al., 2016). Pangenome analysis revealed that functional saturation across the gene coding potential of these species may drive the maintenance and acquisition of multiple strains (Oh et al., 2016). Within the gut, studies have shown that a community of Clostridium species can act synergistically to enhance an immunologic effect greater than any individual species could alone (Atarashi et al., 2013). Similar studies are needed to demonstrate the possible functional advantages of heterogeneous skin strain communities.

1.5 Cutaneous immune microbe dialogue

The skin's immune system and that of other mucosal sites has evolved closely with resident microbes to allow maintenance of commensal partners and elimination of unwelcome transients. The skin accomplishes this task with a sophisticated system of immune surveillance composed of epithelial cells, lymphocytes, and antigen presenting cells in the epidermis and dermis (Nakatsuji et al., 2013). More specifically, within the dermis, there are innate cells (such as macrophages, dendritic cells, and mast cells), innate lymphoid cells (ILCs, including group 2 ILCs and gamma delta T cells), and many adaptive resident lymphocytes (including CD4+ and CD8+ T cells) (Pasparakis et al., 2014; Tong et al., 2015a).

To operate optimally, the skin microbiota, epithelial cells, and both arms of the immune system need to communicate effectively. Keratinocytes can begin this dialogue by sampling microbes on the skin surface via pattern recognition receptors (PRRs) such as Toll-like receptors (TLRs), mannose, and Nucleotide oligomerization domain **(**NOD)-like receptors (Grice and Segre, 2011). These receptors recognize pathogen associated molecular patterns (PAMPs) such as flagellin, nucleic acids, and lipopolysaccharides from bacteria, and mannan and zymosin from fungi. Binding of PAMPs to PRRs triggers innate immune responses resulting in the secretion of antimicrobial peptide (AMPs), cytokines, and chemokines. AMPs, molecules that can rapidly kill and inactivate a diverse range of organisms including fungi, bacteria, and parasites, are our body's first line of defense against pathogens (Gallo and Hooper, 2012). While some AMPs are constitutively expressed, the expression of others can be controlled by members of the skin microbiome, including *P. acnes* (Nagy et al., 2006) and *S. epidermidis* (Naik et al., 2015). However, because microbes can be resistant to AMPs (Cullen et al., 2015; Joo et al., 2016), how these molecules ultimately shape microbial communities is poorly understood.

Studies comparing conventional to germ-free mice showed that microbes are essential for the development of gut-associated immune cells (Lee and Mazmanian, 2010), but the overall structure or seeding of skin-directed immune cells occurs even without microbial colonization (Belkaid and Hand, 2014; Naik et al., 2012). Which is not to say that microbes don't educate the skin immune cells: commensal organisms are essential for proper education of the immune system in responses to pathogens and commensals. In the skin, initial microbial exposure is dependent on delivery mode: vaginally delivered babies first acquire microbes from their mother's vagina, while babies born via Caesarean section acquire microbes from the skin (Mueller et al., 2015). During this postnatal period, the immune system is immature enough to allow microbial colonization in the absence of inflammatory responses (PrabhuDas et al., 2011). This tolerance is dependent on T regulatory (T_{reg}) cells; a subset that have been shown in mice to populate neonate skin, post morphogenesis of the hair follicle when microorganisms are beginning to colonize the site (Scharschmidt et al., 2015). This likely represents a mechanism by which regulatory responses are induced to limit aberrant responses against commensals. Indeed, association of *S. epidermidis* to neonate but not adult murine skin induced *S. epidermidis*-specific FOXP3$^+$ T_{reg} cells that limited inflammatory responses to the skin commensal upon future tissue damage (Scharschmidt et al., 2015). Given the many resident commensals on the skin, it is unsurprising that this site contains one of the highest frequencies of FOXP3$^+$ T_{reg} cells in the body (Belkaid et al., 2002; Suffia et al., 2006). How these regulatory responses are initiated and maintained later in life despite shifts in microbial communities remains poorly understood.

After this initial tolerogenic period, different microbes have been shown to illicit distinct effects on the innate and consequently adaptive immune systems. To date, immune responses induced by the ubiquitous skin commensal *S. epidermidis* have been well described in murine models. First, it was discovered that lipoteichoic acid from *S. epidermidis* cell walls binding TLR2 was sufficient to inhibit inflammatory responses which limited tissue damage and promoted wound healing (Lai et al., 2009). More recently, topically applied *S. epidermidis* was shown to induce increased levels of the proinflammatory cytokine interleukin 1 (IL-1)(Naik et al., 2015; Naik et al., 2012). Expressed by a large numbers of skin cells including keratinocytes, IL-1 is involved in the initiation and amplification of immune responses (Pasparakis et al., 2014). In this particular case, IL-1 promoted skin homing T-cells to produce the cytokines IL-17 and interferon gamma (IFNγ), cytokines important for host defense and inflammatory diseases (Naik et al., 2015; Naik et al., 2012). This particular effect was dependent on the cooperation of 2 of the 4 skin resident dendritic cell (DC) subsets, CD11b^{+} and CD103^{+} (Naik et al., 2015). Chapter 4 explores how different effector subsets are induced depending on the strain of *Staphylococcus epidermidis* that is applied.

Notably, this induction of effector T cells occurred in the absence of classical inflammation in a process termed "homeostatic immunity" (Belkaid and Tamoutounour, 2016). This process represents an essential mechanism whereby different commensals can educate distinct aspects of the immune system to respond to future pathogen exposures. In other words, immune responses to pathogen exposures occur in the context of broader recall responses to diverse microbial antigens (Hand et al., 2012). This is consistent with the skin harboring ~20 million effector lymphocytes (Clark et al., 2006), many of which are likely specific to skin commensals (Belkaid and Tamoutounour, 2016). This concept is demonstrated when mice pre-associated with *S. epidermidis* were better protected against skin infections with *Candida albicans* and *Leishmania major* (Naik et al., 2015; Naik et al., 2012).

Distinctly, when *S. epidermidis* was first introduced via intradermal injection, instead of topically to the mouse, classical inflammatory responses as characterized by infiltrating monocytes and neutrophils were observed alongside IFNγ producing T effector cells (Naik et al., 2015) [Figure 1.5]. This dichotomy highlights how immune responses are mounted in a context dependent manner, meaning immune responses are tailored not only to the identity of the microbe but also their location of detection. Such contextual responses are essential considering *S. epidermidis* typically inhabits the skin as a beneficial commensal but can be a deadly pathogen when in the blood stream (Otto, 2009).

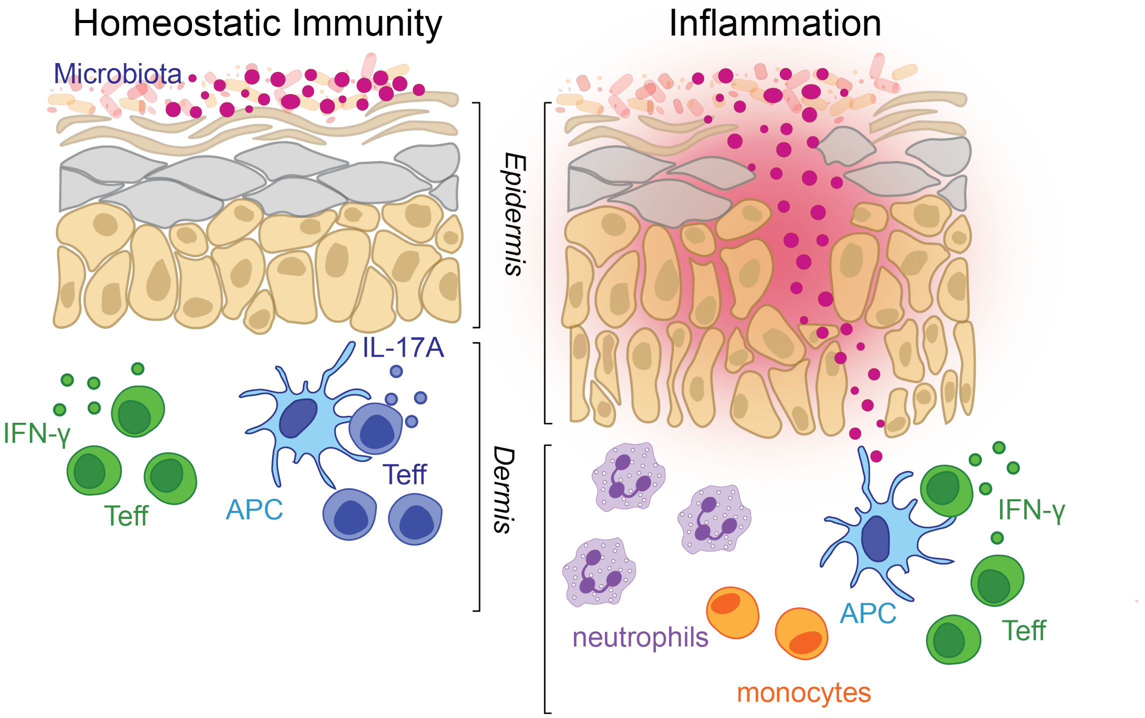


Figure 1.5. Cutaneous immune responses are context dependent.

Immune responses to skin microbes (commensals) vary depending on the localization of the microbe. If a microbe is sensed in the epidermis of the skin, adaptive immune responses develop in the absence of inflammation, a process termed "homeostatic immunity". By contrast, in the circumstance of a barrier breach where a microbe can enter the dermis of a skin, adaptive immune responses develop in the presence of classical inflammation, as defined by the presence of neutrophils and monocytes. This highlights how cutaneous immune responses are compartmentalized depending on the route of microbial exposure.

In addition to effector responses, microbes can also induce regulatory responses. For example, lysates from *Vitreoscilla filiformis*, a bacteria originally isolated from thermal spa water, when applied to mouse skin promote cutaneous T_{reg} accumulation that inhibits T cell proliferation during eczematous-like inflammation (Volz et al., 2014). In addition, secreted products from *S. epidermidis* were shown to promote the tolerance inducing cytokine IL-10 from human DCs *in vitro* (Laborel-Preneron et al., 2015). Identifying additional microbes that induce tolerogenic effects offers many therapeutic opportunities as aberrant immune responses are characteristic of many skin diseases (Belkaid and Tamoutounour, 2016).

As mentioned, several studies exist demonstrating that microbes can have distinct effects on the immune system. Now, future studies are needed to explore the microbial molecules/products that are mediating these responses and how the immune system is sensing their presence. In addition, immunologic tools should be developed to track these commensal specific immune responses (Newell and Davis, 2014). Such tools would allow visualizing these cells in the tissue, tracking their persistence overtime, and seeing how they respond to pathogens. Understanding these details is necessary to transition from observations to therapeutics for blocking undesired effects or inducing the desired ones.

1.6 Microbial roles in skin inflammatory disorders

In addition to educating our immune system, microbes play the essential role of inhibiting colonization of harmful bacteria in a process termed colonization resistance (Buffie and Pamer, 2013). However, in certain contexts, normally beneficial bacteria can be implicated in disease. In fact, many common skin diseases are associated with an altered microbial state, termed dysbiosis (Iebba et al., 2016). This dysbiosis is often driven by common commensal species. For example, the prevalent teenage malady acne vulgaris is a chronic inflammatory skin condition associated with the bacteria *Propionibacterium acnes* (Leyden et al., 1975), the most abundant organism in the microbiome of healthy adults (Fitz-Gibbon et al., 2013; Tomida et al., 2013). The reality that almost all adults are colonized with *P. acnes* but only a minority have acne highlights the importance of studying diseases in the broader context of host genetics, immune or barrier defects, microbiome, and the environment. For example, beside *P. acnes* presence, increased sebum secretion is associated with the pathophysiology of acne as secretion rates correlate well with severity of clinical manifestations (Picardo et al., 2009). Other common skin diseases associated with an obvious dysbiosis are athletes foot and fungal outgrowth, seborrheic dermatitis and Malassezia (Gaitanis et al., 2012), and atopic dermatitis and *Staphylococcus aureus* blooms (Leyden et al., 1974).

Atopic dermatitis, or eczema, is an extremely heterogeneous disease with multiple contributing factors including epidermal barrier impairment, type 2 immunity, and skin microbes. In addition to *S. aureus* being commonly cultured from AD skin (Leyden et al., 1974), there are additional factors that support the microbiome playing an influential role. 1) AD is clinically treated with combinations of antimicrobial approaches (e.g. antibiotics and dilute bleach baths) and anti-inflammatory or immunosuppressive medications (Huang et al., 2009). Their success correlates with decreases in Staphylococcal relative abundance (Kong et al., 2012). 2) AD flares commonly manifest at the bend of the elbow and the back of the knee two sites that are physiologically classified as moist and have shared microbial communities (Grice et al., 2009). 3) The majority of children will outgrow AD prior to puberty, a time when increased levels of hormones stimulate sebaceous glands to produce additional sebum thus favoring the expansion of lipophilic bacteria, such as Propionibacterium and Corynebacterium (Oh et al., 2012).

In a longitudinal study of AD patients, 16S rRNA sequencing of clinical samples showed that the relative abundance of Staphylococcal species, particularly *S. aureus* and *S. epidermidis,* increased in flare versus post flare state and the abundance of this Staphylococcus correlated with more severe disease. Chapter 5 presents a higher resolution longitudinal study of AD flares based on whole genome metagenomic samples. Specifically we address whether different species of Staphylococcus exist as homo- or heterogeneous communities of strains and how those strain communities respond throughout the disease course. In a different paper, comparing the baseline skin microbiome of adult AD patients and controls, the authors identified microbial signatures enriched for Streptococcus and Gemella but depleted for Dermacoccus in AD-prone individuals (Chng et al., 2016). At a functional level, they show that the AD-prone microbiome is primed to generate excess ammonia, providing a microbial explanation for the high pH levels observed during AD flares.

Because of *S. aureus*'s association with AD, other skin disease, and also blood stream infections, many studies have focused on interactions between *S. aureus*, its toxins, and the immune system. For example *S. aureus* produced δ-toxin induces degranulation of mast cells, which promotes both innate and adaptive type 2 immune responses (Nakamura et al., 2013). *S. aureus* α-toxin can also induce IL-1β production from monocytes that may consequently promote a Th17 response, or $CD4^+$ cells making the cytokine IL-17 (Niebuhr et al., 2011). In addition to effecting classical immune cells, *S. aureus* has also been shown to trigger adipocytes to rapidly proliferate and to produce increased level of the AMP cathelicidin as a host defense mechanism (Zhang et al., 2015). These example demonstrate the many ways *S. aureus* could initiate or amplify skin disorders in the broader context of barrier defects or altered immunity. In fact, it has been demonstrated that in the context of barrier breach, *S. aureus* is able to breach the epidermis into the dermis where it encounters immune cells and triggers expression of the inflammatory cytokines, IL-4, IL-13, IL-22, and TSLP (Nakatsuji et al., 2016). Chapter 5 highlights how *S. aureus* can induce inflammatory responses in the absence of barrier breach and in a strain specific manner.

Although the inflammatory potential of *S. aureus* has been demonstrated and dysbiosis is common to many skin diseases, it is still unknown whether these microbial changes are a consequence of the disease, resulting from the release of extracellular matrix proteins with the itch-scratch of AD, or whether *S. aureus* contributes as an initiator of the disease. We can begin to differentiate these two models by comparing skin microbial communities in mice with various skin barrier or immunologic defects. For example, mice with mutations in matriptase, a serine protease essential for proper skin integrity, exhibit flaky skin and increased expression of antimicrobial peptides (Scharschmidt et al., 2009). In a separate study, mice deficient in disintegrin, a metalloproteinase domain-containing protein (Adam17), also experienced eczematous dermatitis from a defective barrier and microbial dysbiosis (Kobayashi et al., 2015). In this case, inflamed skin was characterized by an overgrowth of *Corynebacterium mastidis*, *Corynebacterium bovis*, and *S. aureus*. Targeted antibiotic treatment of these animals was sufficient to reverse the dysbiosis and eliminate skin inflammation.

Studying human primary immunodeficiency patients (PID) provides an opportunity to observe the influence of altered immunity on microbial communities. To study this, skin microbiome samples were taken from patients with the rare monogenic primary immune deficiencies. Compared to healthy individuals, the skin of PID patients is more ecologically permissive with decreased site specificity and temporal stability, and colonization with opportunistic fungi (Candida and Aspergillus) and bacterial species absent in controls, including Clostridium species and *Serratia marcescens (Oh et al., 2013)*. Despite this increased permissiveness, the new species belong within phyla commonly associated with the skin, i.e. Firmicutes and Proteobacteria. This implies that organisms outside these primary phyla are perhaps unable to stably survive in the nutrient poor environments of the skin. Whole genome metagenomic sequencing of samples from these patients are needed to observe fluctuations in the viral communities. Such studies would be interesting considering PID patients, particularly those with mutations in dedicator of cytokinesis protein 8 (DOCK8), commonly suffer from viral skin infections (Chu et al., 2012).

Overall, this thesis will provide previously unexplored resolution of human skin microbial communities. Chapter 2 will explain the pipelines developed to analyze whole genome metagenomic datasets at strain-level resolution. Chapter 3 will highlight the multi-kingdom microbial communities present in healthy adults over time with special emphasis on strains of *P. acnes* and *S. epidermidis*. Chapter 4 will discuss the functional potential of the heterogeneous *P. acnes* and *S. epidermidis* strain communities. Finally, Chapter 5 will present results from a longitudinal study of atopic dermatitis patients along with functional data linking microbes with the disease.

CHAPTER 2 Accurate identification of microbes in unassembled sequencing data with Clinical Pathocope

2.1 Abstract

The use of sequencing technologies to investigate the microbiome of a sample can positively impact patient healthcare by providing therapeutic targets for personalized disease treatment. However, these samples contain genomic sequences from various sources that complicate the identification of pathogens. Here we present Clinical PathoScope, a pipeline to rapidly and accurately remove host contamination, isolate microbial reads, and identify potential disease-causing pathogens. We have accomplished three essential tasks in the development of Clinical PathoScope. First, we developed an optimized framework for pathogen identification using a computational subtraction methodology in concordance with read trimming and ambiguous read reassignment. Second, we have demonstrated the ability of our approach to identify multiple pathogens in a single clinical sample, accurately identify pathogens at the subspecies level, and determine the nearest phylogenetic neighbor of novel or highly mutated pathogens using real clinical sequencing data. Finally, we have shown that Clinical PathoScope outperforms previously published pathogen identification methods with regard to computational speed, sensitivity, and specificity. Clinical PathoScope is the only pathogen identification method currently available that can identify multiple pathogens from mixed samples and distinguish between very closely related species with very little coverage of the genome. Furthermore, Clinical PathoScope does not rely on genome assembly and thus can more rapidly complete the analysis of a clinical sample when

compared with current assembly-based methods. Clinical PathoScope is freely available at: http://sourceforge.net/projects/pathoscope/.

Note: The work presented in this chapter has been previously published in (Byrd*, Perez-Rogers* et al., *BMC Bioinformatics* 2014).

2.2 Introduction

Despite recent advances in diagnostic and preventative medicine, infectious diseases still account for a large proportion of the disease burden and mortality worldwide, particularly in low-income areas and developing countries (WHO, 2004). Current clinical diagnostic tests for identifying infection-causing pathogens utilize limited technologies such as polymerase chain reactions (PCR), Sanger sequencing, or cell culture. These methods typically focus on identifying only a single pathogen at a time and often lack the specificity required to distinguish between closely related species or strains of the same species. Bacterial cultures can accurately identify culturable pathogens, but usually require 4-5 days to complete and cannot be conducted for all pathogens (Didelot et al., 2012). Microarray technologies, such as the Virochip (Chen et al., 2011a), have been shown to be useful in the space of pathogen identification. Microarrays, such as these, are designed to detect both known and novel pathogens through the use of high-sensitivity probes and probes that map to conserved genomic regions. While useful for broad spectrum screening of clinical samples, this technology is limited in that probes must be continually designed and updated to support the ever growing number of genomic sequences in public databases.

In recent years, researchers have taken advantage of innovations in sequencing technologies to more rapidly identify and characterize pathogens responsible for disease outbreaks, including the West Nile Virus (Lanciotti et al., 1999), H1N1 influenza (Deng et al., 2011; Greninger et al., 2010; Kuroda et al., 2010), cholera (Chin et al., 2011), Escherichia coli (Frank et al., 2011; Rasko et al., 2011; Rohde et al., 2011; Turner, 2011), *Salmonella* (Lienau et al., 2011), and antibiotic resistant *Klebsiella pneumonia* (Snitkin et al., 2012). Traditionally, sequencing a single sample has taken as long as several days or weeks using the most common platforms. Recent commercial efforts, however, have reduced this time to a few hours or days (Rothberg et al., 2011). Within the next few years, newer technologies are promising sequencing runs in less than an hour with a cost of under one hundred dollars (Rothberg et al., 2011). Once these technologies become widely accessible, the use of sequencing as a diagnostic tool in the clinic will have great potential for more personalized medical applications. The rapid and accurate analysis of next-generation sequencing data, however, remains a challenge for many reasons. The sheer volume of data, for example, is difficult to deal with computationally without significant computational resources (e.g., a typical sequencing run on the Illumina HiSeq 2500 can yield 300M million reads requiring 30 GB of storage capacity and significant RAM requirements for processing). Furthermore, DNA from host genomes or commensal species will often dominate clinical samples and sequencing error can swamp out diagnostic signal. These challenges highlight the need for the development of highly sensitive algorithms that can distinguish among closely related pathogenic strains in a computationally efficient manner.

Current sequencing-based diagnostic methods (Bhaduri et al., 2012; Brady and Salzberg, 2009; Huson et al., 2007; Kostic et al., 2011; Naeem et al., 2013; Patil et al., 2011; Segata et al., 2012) require thousands of reads from the pathogen and include computationally intensive steps such as genome assembly, multiple genome alignments, extensive homology searches, and/or phylogeny estimation, with some methods taking upwards of three days to complete a single run (Kostic et al., 2011). Additionally, these methods fail to accurately identify pathogens at the strain level and will often assign ambiguously aligned reads to higher taxonomic levels which may lead to a nonspecific or incorrect diagnosis and the administration of ineffective clinical treatments. Such was the case during the European outbreak of hemorrhagic *Escherichia coli*, which resulted in 3,800 infections and 54 deaths across 13 countries due to a 3-week delay in appropriate intervention (Frank et al., 2011). The challenges encountered when diagnosing viral and bacterial pathogens in the clinic reinforce the need for a streamlined sequencing protocol and a highly sensitive computational method by which strain specific identification can be rapidly achieved. By helping clinicians to direct treatment and avoid misdiagnoses, the identification of viral and bacterial pathogens in clinical samples will directly benefit patients suffering from a variety of infectious diseases (Bibby, 2013). In particular, assigning a viral rather than bacterial cause to an infection may help alleviate the antibiotic overuse that is common in clinical practice today (Wylie et al., 2012). Recent editorials and reviews express concern that analysis, rather than data generation, is likely to be the limiting factor for sequence-based clinical pathology; thus, clearly highlighting the need for ‘clinic-ready’ software tools and approaches (Chan et al., 2012; Didelot et al., 2012; Dunne et al., 2012; Torok and Peacock, 2012; Walker and Beatson, 2012).

Here we present Clinical PathoScope, a rapid alignment and filtration pipeline for accurate viral and bacterial pathogen identification using unassembled sequencing data. Using a variety of clinical samples and simulated scenarios, we demonstrate our method's ability to differentiate between pathogens, identify multiple pathogens in a single clinical sample, and identify the closest relative to highly mutated and novel strains. Clinical PathoScope builds on the previous success of PathoScope v1.0 (Francis et al., 2013), which capitalizes on a Bayesian statistical framework to process an alignment file and provide posterior probability profiles of organisms present. While PathoScope v1.0 showed success when used with purified samples, it was necessary to develop a method to remove potential contaminating sequences from the host and commensal microbes for host-dominated clinical samples. Clinical PathoScope incorporates the original PathoScope algorithm into a novel pipeline that allows users to go directly from metagenomic sequencing reads to a list of organisms present in a sample in one easy step and in a clinically relevant timeframe. For convenience, we provide bacterial and viral databases curated from NCBI; however, custom databases can easily be incorporated as well. Taken together, these features make Clinical PathoScope the fastest and most accurate pipeline currently in the literature for identifying strain-specific pathogens in clinical samples without the need for genome assembly. Clinical PathoScope (version 1.0) is freely available at: http://sourceforge.net/projects/pathoscope/.

2.3 Methods

In order to develop the Clinical PathoScope framework, we have accomplished the following essential tasks for pathogen identification in clinical samples: 1) selection of the most appropriate alignment algorithm and parameters for optimal performance on clinical samples, 2) evaluation of filtering approaches to efficiently remove reads from a clinical sample that originated from host, non-target, or non-pathogenic genomes, and 3) the evaluation and comparison of Clinical PathoScope with existing approaches using multiple real datasets [Figure 2.1]. Details regarding the specific methods evaluated, pipeline modules, and results observed are given in the subsequent sections. Finally, we have implemented these results into a highly sensitive and efficient pipeline that is user-friendly and approachable by physicians and researchers without the requirement of advanced computational expertise.

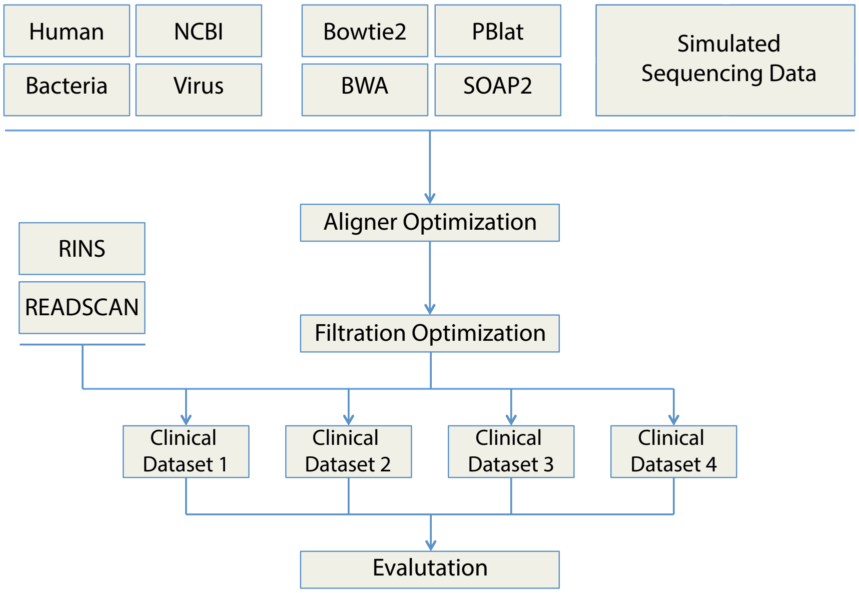


Figure 2.1. Workflow employed to develop the Clinical PathoScope pipeline.

Three reference genome libraries were downloaded from NCBI. Four alignment algorithms were tested and evaluated on five simulated clinical sequencing samples. Each aligner was parameter tuned and optimized and Bowtie2 was selected as the choice aligner for the Clinical PathoScope pipeline. The order with which reads are aligned to the reference libraries was determined and the performance of Clinical PathoScope was evaluated using four clinical datasets. Furthermore, we compared our results against those produced by existing technologies.

2.3.1 Clinical PathoScope pipeline development & evaluation

The Clinical PathoScope pipeline consists of three primary steps: 1) optimized read alignment, 2) host and non-target genome filtration, and 3) ambiguous read reassignment. We developed the optimized Clinical PathoScope algorithm using a set of simulated clinical samples (described below) and later validated our method and compared our results against existing approaches using multiple clinical datasets, some of which are original to this publication.

2.3.2 Reference genome library curation and processing

One of the most important steps for the accurate identification of benign and pathogenic genomes is to build a comprehensive genome library containing all species and strains likely to be present in the sample. This is a critical step as Clinical PathoScope can only identify organisms or their nearest neighbors if they are present in the library. In order to maximize the characterization of all reads within a given clinical sample, our method aligns reads against three broad categories of reference genomes. The human host library consisted of two sequences totaling 3.2 gigabase-pairs (Gbps); the GRCh37/hg19 build of the human genome, as well as the human ribosomal DNA sequence [GenBank:U13369]. The ribosomal reference was included in order to remove several false positive alignments to viral genomes that share sequence similarity with human ribosomal RNA (gi|401829614|Shamonda virus segment L, gi|109255272|Choristoneura occidentalis granulovirus, and gi|401829625|Simbu virus segment L). The bacterial library was downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz, 12/15/12) and contained 2,402 complete reference genomes and 1,759 plasmid sequences. In all, this bacterial library consisted of 7.7 Gbps of DNA sequence. Due to restrictions enforced by some of the aligners with regard to index size, it was necessary to split this library into two smaller segments to facilitate proper alignment. Finally, the viral library was also obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/all.fna.tar.gz, 1/10/13). For genomes in which multiple segments were available, all segments for a given genome were concatenated into a single contiguous sequence with each segment separated by a series of null characters (N's). In total, the viral library contained 3,738 complete genomes and 110 megabase-pairs (Mbps) of total sequence.

2.3.3 Generation of simulation study datasets

We simulated two sets of five *in silico* clinical samples to represent a variety of clinical scenarios including infections with two or more disease causing and benign pathogens, infections with a pathogen having closely related substrains (e.g. Human adenovirus), and infections with highly mutated pathogens. The first set of simulated samples was used to evaluate several alignment algorithms and to optimize the architecture of the Clinical PathoScope pipeline. The second set was then used to evaluate the efficacy of Clinical PathoScope alongside existing technologies. Importance was placed on implementing accurate mutation rates, genome diversity, and relative compositions. Functioning as positive controls, these data were essential to develop a robust pipeline for pathogen identification. Each sample was composed of human, bacterial, and viral sequences mimicking the microbiota found in sequencing data from nasopharyngeal samples during a respiratory tract infection (Bogaert et al., 2011; Yang et al., 2011). Specifically, 10 million 100-base reads were generated for each sample with 90% of reads originating from the host transcriptome (human RNA), 9% from bacterial genomes, and 1% from viral genomes. The first set of simulated samples contained sequencing reads from five bacterial and six viral genomes at various depths of coverage. This was essential to determine how each aligner and pipeline architecture performed with respect to the number of reads originating from each genome. The second set of simulated samples was designed as a more challenging and realistic dataset and was used to evaluate our optimized approach. Each sample contained sequences from six viral genomes and twenty-five bacterial genomes. The number of reads originating from each viral genome ranged from 10 to 63,640. To determine a realistic bacterial landscape for these samples, we downloaded and aligned three anterior nares samples [SRA: SRS011105, SRS012291, SRS013637] from the Human Microbiome Project (http://hmpdacc.org/HMASM/) and selected 25 of the most common bacterial strains (19 unique species) to be included in our simulation. The number of reads originating from each bacterial genome was determined by sampling a Gaussian distribution such that the number of bacterial reads per sample totaled 900,000. Reference genomes for each of the representative species were obtained from NCBI's RefSeq database (Pruitt et al., 2012) and samples were simulated using the next-generation read simulator, Mason (Holtgrewe, 2010), employing its ‘Illumina sequencing’ error-model (Mason illumina -s $seed -N

$numReads -sq -n $readLength -i -hs $snpRate -hi $indelRate -hnN -nN -o $outputFile.fastq $refGenomes.fa). Previously published species or kingdom specific mutation rates for SNPs and indels were applied to the human (Genomes Project et al., 2012), bacterial (Chen et al., 2009), and viral (Sanjuan et al., 2010) genomes to accurately capture the variability inherent in clinical samples. The simulated datasets are available for download on the PathoScope software distribution site and will be useful for benchmarking and comparing future metagenomic analysis pipelines.

2.3.4 Alignment optimization

We evaluated and compared four publicly available alignment algorithms (Bowtie2.0.0 (Langmead and Salzberg, 2012), BWA 0.6.2 (Li and Durbin, 2009), PBLAT 2.0.0 (Kent, 2002), SOAP2 2.21 (Li et al., 2009)) based on three criteria, namely, 1) run time, 2) sensitivity and 3) specificity by aligning our first set of five simulated samples against the human, bacterial, and viral reference libraries described above [Figure 2.2]. Run time was measured as cpu minutes using 8 cores and a single 2.3 GHz AMD Opteron processor on the Boston University Medical Campus LinGA cluster. Using the resulting alignment files and the known origin of the reads, sensitivity was measured as the number of true positives divided by the number of true positives plus false negatives, and specificity was measured as the number of true negatives divided by the number of true negatives plus false positives. Our goal was to identify the algorithm and parameters that provided the best balance of our three evaluation criteria. Additionally, we examined the effect of varying the length of each read on the number of reads correctly aligned to the reference genomes using the first 25, 50, 75, and 100 basepairs, as well as the full-length sequence. Evaluating variable read lengths served multiple purposes: 1) determining whether aligning the entire read was necessary, or if aligning a smaller segment of the read performed just as well, 2) identifying optimal sequence read size for future studies, and 3) evaluating whether aligning a smaller portion of the read can replace the need for a computationally intensive spliced-read alignment algorithm for reads from host/filter genomes that contain spliced gene transcripts.

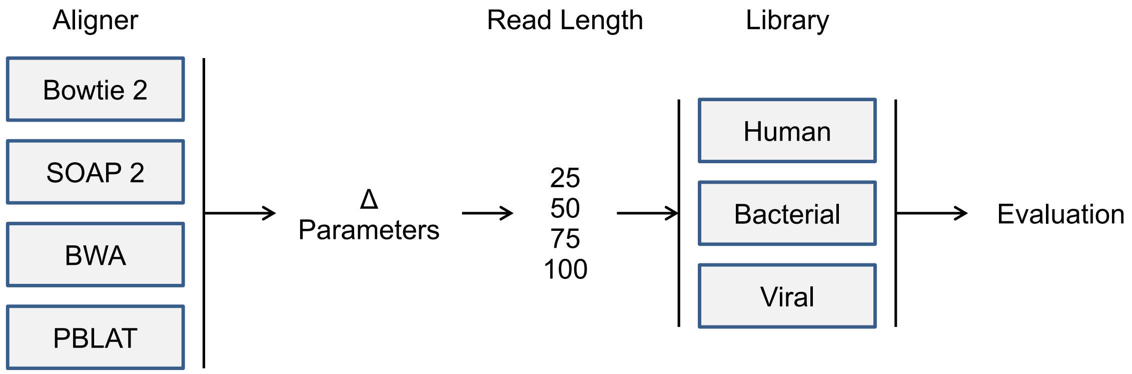


Figure 2.2. Alignment optimization variables and methods.

The internal parameters for each of the four aligners were varied and tuned. Additionally, the length of each read aligned was varied. For each unique aligner-parameter-read length configuration, the sensitivity, specificity, and run time when aligning the simulated samples against the reference genome libraries was calculated.

2.3.5 Filtration optimization

We employed a computational subtraction methodology (Xu et al., 2003) in which reads are sequentially aligned against a series of reference genomes to determine their origin. For our purposes, we aligned reads against libraries of reference genomes originating from human, bacteria, or viruses. Within our pipeline, reads that align to the target library (e.g. viral library for virus detection) are retained while reads that align to the host (e.g. human library) and non-target (e.g. bacterial library) sequences are removed. The effects of varying the order of subtraction were examined by comparing the resulting alignment sensitivity, specificity, and pipeline run time using all six permutations of our three libraries. Additionally, we evaluated the effect of using the PathoScope expectation maximization (EM) algorithm (Francis et al., 2013) to minimize false positive mappings by reassigning reads with ambiguous alignments to their correct genome of origin. A detailed diagram of the overall experimental design is shown in Figure 2.1. The subtraction methods evaluated for use in our pipeline as well as the optimal method are shown in Figure 2.3.

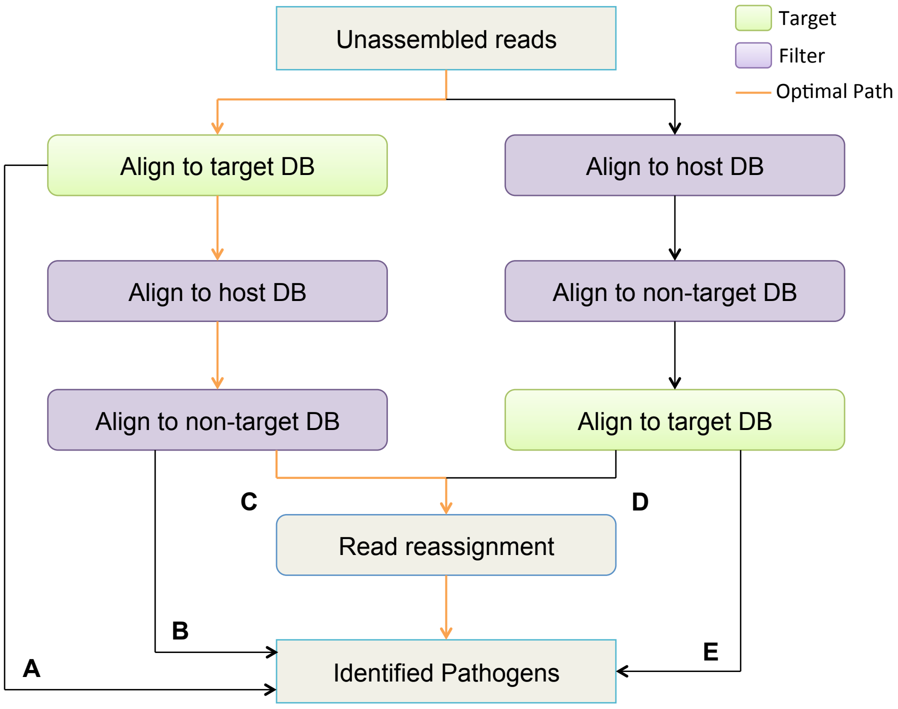


Figure 2.3. Subtraction and filtration optimization methods.

Various filtration methods were tested in an effort to minimize computational burden and maximize accuracy. Approaches tested include A) Naïve Approach, B) Target Centric, C) Target Centric + Reassignment, D) Host Centric + Reassignment, and E) Host Centric. Post filtration, all reads are aligned against the target genome library. The resulting read alignments are reassigned to the correct genome of origin using the PathoScope Expectation Maximization algorithm.

2.3.6 Clinical datasets

Prostate Cancer Cell Line (PCCL): The PCCL dataset (Prensner et al., 2011) has been leveraged in previous studies as a positive control and a means for comparing algorithm run time. This dataset is derived from a prostate cancer cell line infected with the human papillomavirus serotype 18. The RNA sequencing was performed using an Illumina GA II sequencer and 26,958,682 reads (40 bases each) were publically available [SRA:SRR073726].

New World Titi Monkey Adenovirus Outbreak (TMAdv): Sequencing reads from two New World titi monkeys (*Callicebus cupreus*) infected with a highly divergent adenovirus (Chen et al., 2011b) make up the third dataset used to evaluate Clinical PathoScope. The samples originated from an outbreak of an unknown virus in a colony of titi monkeys in California. Tissue samples were obtained from the lungs of two titi monkeys during necropsy and were sequenced together using the Illumina GA IIx for 73 cycles in both directions yielding 12,393,506 reads (73 bases). Chen *et al*. identified the cause to be a new highly divergent species of adenovirus that was subsequently assembled and so named Titi Monkey adenovirus (TMAdv). We supplemented our host library with the most closely related, fully sequenced simian species, *Callithrix jacchus* [GenBank:PRJNA46205]. As a positive control, we included the TMAdv genome in our target library and validated that Clinical PathoScope accurately distinguished the TMAdv from all other adenovirus genomes.

Tuberculosis in a Mummy: Sequencing reads from a 200 year old mummy infected with tuberculosis were obtained from a previous study (Chan et al., 2013) and used to evaluate Clinical PathoScope's ability to detect bacterial pathogens. The sample was collected from lung tissue taken from the left side of the thorax of a mummified body. Pulmonary tuberculosis was suspected because of the cathectic state of the body and confirmed based on PCR analyses. As further validation, the sample was sequenced on the Illumina Miseq instrument for 300 cycles in both directions yielding 5,541,400 reads with an average length of 297 basepairs; the reads were retrieved from Sequence Read Archive with accession number SRP018736. For analysis with Clinical PathoScope, the reads were split into 12,261,862 reads of approximately 100 bases in length.

16S Amplimer Sequencing (16S): In addition to testing our approach on *in silico* and previously published clinical datasets, we validated our approach on data from our own clinical samples. Under an IRB-approved protocol, deep endobronchial aspirates from 3 patients intubated for mechanical ventilation were obtained after the aspirate had been used for microbiologic testing directed by their medical team. The bacteriological staining of aspirate samples revealed the presence of Gram-negative bacteria, and bacterial culture from aspirates identified abundant *Pseudomonas* (patients F1 and G1) and *Enterobacter* (patient H1), with opportunistic flora in all samples. All three patients were on antibiotic treatment regimen prior to the collection of samples. Patient F1 was treated with a combination of aminoglycoside (gentamicin and tobramycin) and polymyxin (colistin) antibiotics; patient G1 was on gentamicin/tobramycin regimen only, and patient H1 was treated with third generation cephalosporin antibiotics (ceftazidime). In addition to clinical samples, we collected the bacterial DNA from gram-positive and gram-negative ATCC reference strains: *Staphylococcus aureus* (ATCC No. 25923 - MSSA), *Enterococcus faecalis* (ATCC No. 51299), *Pseudomonas aeruginosa* (ATCC No. 27853), *Escherichia coli* (ATCC No. 25922). Total DNA from these samples was isolated by centrifugation, and solubilization of the pellet using the Sigma GeneElute kit combined with a lysis buffer by mixing together the Gram+ and Gram- buffers supplemented with lysozyme (2.115X10^6 units/mL), lysostaphin (200 units/mL), mutanolysin (5000 units/mL). Nanodrop and Qubit measurement of concentrations were used to quantify DNA. After DNA isolation, we amplified the 16S rRNA using the U1492R, Tm 49.44 (GGTTACCTTGTTACGACTT) and B27F, Tm 41.67 (AGAGTTTGATCCTGGCTCAG) universal primers using 800 ng of template. The amplimers were ligated into SMRTbells and sequenced on a Pacific Biosystems RS. The sequencing yielded an average of 4,127 reads per sample, averaging 1,178 bases long. For analysis with Clinical PathoScope, the PacBio reads from each sample were split into 100 base segments that were then treated as individual reads, generating on average 39,183 reads of 100 bases per sample. To accommodate the high homologies of 16S RNA sequences from different bacterial species and strains, the alignment parameters for this dataset were tightened compared to the viral samples, allowing 1 mismatch per 100 bases during alignment, and allowing for multiple 'best' hits per read (e.g. Bowtie2 'k' set at 1,000). These data were submitted to the NCBI Sequence Read Archive (SRA) database under accession number SRP028704.

2.3.7 16S Phylogenetic inference

We took all genomes from GenBank's RefSeq database belonging to *Pseudomonas*, *Enterobacter*, and *Acinetobacter* genera (56 taxa) and generated a BLAST database, which we queried with a full-length 16S rDNA sequence (Altschul et al., 1990). We selected one copy per species and aligned the resulting dataset using a secondary structure aware algorithm (Q-INS-i) as implemented in MAFFT (Katoh et al., 2005). We ran 10 independent Maximum Likelihood searches in RAxML (Stamatakis, 2006) (1000 bootstraps) assuming a GTR nucleotide substitution model with gamma distributed rate heterogeneity. Additionally, we obtained diagnostic characters defining particular species using the phylogeny-aware algorithm implemented in CAOS (Sarkar et al., 2008).

2.3.8 Clinical dataset preprocessing

The four clinical datasets were used to evaluate our Clinical PathoScope pipeline and to compare our method against previously published algorithms. A summary of these datasets is shown in Table 2.1. Overview of clinical datasets used to evaluate Clinical PathoScope. Extensive quality control was performed uniformly on each of the datasets to remove low quality and artificial sequences using PrinSeq (Schmieder and Edwards, 2011) (-derep 123; -lc_method dust; -lc_threshold 40) and Cutadapt (Martin, 2011), respectively. For each read, bases having a Phred quality score less than 20 were trimmed from the 3' end and reads with a median quality score below 20 were removed. Low complexity and redundant reads were determined using PrinSeq and removed along with adapter and primer sequences. A minimum read length of 25 base pairs was strictly enforced for trimmed reads to facilitate accurate sequence alignment. Reads that failed to meet the length requirement were not considered for further analysis.

Name	Accession	# Samples	Read Length	Avg. # Reads
PCCL	SRR073726	1	40	26958682
CALRTI	Yang et. al	14	36	3907924
TMAdv	SRR167721	1	75	12222012
16S	SRP028704	8	1178	4127

Table 2.1. Overview of clinical datasets used to evaluate Clinical PathoScope.

2.3.9 Comparison to published algorithms

Clinical PathoScope was evaluated alongside two existing pathogen identification algorithms, RINS (Bhaduri et al., 2012) and READSCAN (Naeem et al., 2013) to emphasize the major differences in performance between assembly-based approaches and our implementation of computational subtraction with varying read length and ambiguous read reassignment. All three methods were compared based on their ability to rapidly identify the pathogens present in the clinical datasets described above. We also considered several published metagenomic-like pipelines such as CloVR-Metagenomics (Angiuoli et al., 2011), IMSA (Dimon et al., 2013), LMAT (Ames et al., 2013), and metAMOS (Treangen et al., 2013) in the context of pathogen identification.

2.4 Results and discussion

2.4.1 Comparison of alignment algorithms

The internal parameters for each alignment algorithm were evaluated and tuned to maximize alignment sensitivity and specificity as well as to minimize run time by mapping reads from our first set of simulated samples to the reference libraries. The average alignment results and confidence intervals of each algorithm using optimized parameters and read lengths are shown in Table 2.2. When aligning reads to the human library, SOAP2 was on average 30.5% faster than Bowtie2; however Bowtie2 had a 15.0% higher average sensitivity at 90.2% and a more consistent run time. For alignments to the viral library, PBLAT had the highest average sensitivity of 99.8%. Bowtie2 also achieved a high average sensitivity of 98.1% with an 80% reduction in average runtime compared with PBLAT. For alignments to the bacterial library, PBLAT had the highest average sensitivity of 98.9%; however, it took almost 20 times longer than Bowtie2, which had an average sensitivity of 79.8%. Overall, Bowtie2 offered the best combination of sensitivity, specificity, and speed when aligning reads against the human, bacterial, and viral libraries.

	Human		Virus		Bacteria	
	Time (m)	Sensitivity	Time (m)	Sensitivity	Time (m)	Sensitivity
		Specificity		Specificity		Specificity
Bowtie2	8.2 ± 0.0	90.2 ± 0.0	3.3 ± 0.6	98 .1 ± 0.6	15.8 ± 1.6	79.8 ± 0.1
		100.0 ± 0.0		99.8 ± 0.2		100.0 ± 0.0
BWA	22.8 ± 3.2	89.9 ± 0.0	6.5 ± 1.4	76.8 ± 5.4	-	-
		100.0 ± 0.0		99.8 ± 0.2		-
SOAP2	5.7 ± 1.6	76.7 ± 0.0	3.9 ± 0.8	50.3 ± 5.4	23.3 ± 2.2	27.7 ± 0.0
		100.0 ± 0.0		99.9 ± 0.1		100 ± 0.0
PBLAT	61.2 ± 6.8	78.2 ± 0.0	16.7 ± 1.3	99.8 ± 0.1	306.3 ± 23.3	98.9 ± 0.0
		100.0 ± 0.0		99.6 ± 0.2		52.7 ± 0.0

Table 2.2. Simulation study alignment statistics using optimal model parameters.

Each aligner was used to align the first set of five simulated sequencing samples (10 million 100 base-pair reads) against each of the three genome libraries using optimal parameters. The average run time, sensitivity, and specificity as well as confidence intervals for each alignment are reported. BWA failed to run to completion with the bacterial library.

2.4.2 Impacts of read length

We evaluated the effect of varying the length of each read used during alignment to further maximize the sensitivity, specificity, and minimize run time. Temporary read splitting and trimming allows clinical samples from any sequencing technology to be analyzed without compromising the speed and accuracy of the short read aligner or losing the alignment specificity of longer reads. For the five simulated samples, varying read length had a larger impact on runtime and sensitivity than adjusting internal parameters. Using Bowtie2 as our primary aligner, 10 million 50 base reads were aligned against the human library in an average 28 minutes, while aligning 100 base reads took on average 40 minutes. Depending on the reference library used, increasing read length may or may not increase sensitivity. Bowtie2 aligned 50 base reads to the human library with an average sensitivity of 90% and 100 base reads with a decreased average sensitivity of

75%. This trend can be explained by the splice junctions found in human transcriptome sequences. With fewer bases, the odds of a read spanning a splice junction are smaller and the read will be more likely to align. Conversely, when aligning reads against the bacterial and viral libraries, the average sensitivity is 10-20% higher using 100 base reads compared to 50 base reads. To evaluate if longer reads continue to increase sensitivity, a subset of 150 base simulated bacterial reads were tested. Results indicate that splitting the 150 base reads into 100 base and 50 base segments increased sensitivity by approximately 4 percent compared to leaving the reads at the full length of 150 bases. Thus, upon initiation, Clinical PathoScope splits all long reads into fragments with a maximum length of 100 bases.

2.4.3 Library alignment and filtering order

Various filtration methods were evaluated in an effort to minimize computation burden and maximize accuracy. Five subtraction frameworks were evaluated: A) Naïve Approach, B) Target Centric, C) Target Centric + Reassignment, D) Host Centric + Reassignment, and E) Host Centric [Figure 2.3]. In the target centric approaches, reads are first aligned against the target library followed by the host and non-target libraries. Conversely, in the host centric approaches, reads are first aligned against the host and non-target libraries and then against the target library. The naïve approach, or only aligning to the target library, took the least amount of time, but resulted in the highest number of false positives. While both the target centric and host centric filtration approaches yielded similar results in terms of accuracy, the target centric approaches required ten fewer minutes to run to completion than the host centric approaches. The target centric approaches were more efficient because a greater number of sequences were removed by initially mapping reads to the target library than to the host library, thus reducing computational burden for subsequent alignments. To determine the impact of the read reassignment algorithm, we compared the sensitivity of both target centric approaches by analyzing our second set of simulated samples. With viral pathogens as the target library, the target centric approach with read reassignment achieved an average sensitivity of 97.8% for species and strain level identifications [Figure 2.4]. Without the reassignment algorithm, the target centric approach achieved an average sensitivity of 90.3% and 78.1% at the species and strain level, respectively. Concurrently, with bacterial pathogens as the target library, the target centric method with reassignment achieved an average sensitivity of 77.6% and 72.8% at the species and strain levels [Figure 2.4], respectively, compared with 52.8% and 41.7% for species and strain specific identifications without read reassignment. These dramatic improvements in sensitivity between methods with and without read reassignment demonstrate the necessity of this algorithm within the Clinical PathoScope pipeline. The performance difference between viral and bacterial identification can be directly attributed to the mixture of bacterial pathogens present in these simulated samples. When two very closely related strains of the same species are present in a given sample, Clinical PathoScope will tend to reassign reads which aligned to both strains to the strain with more uniquely identifying sequences.

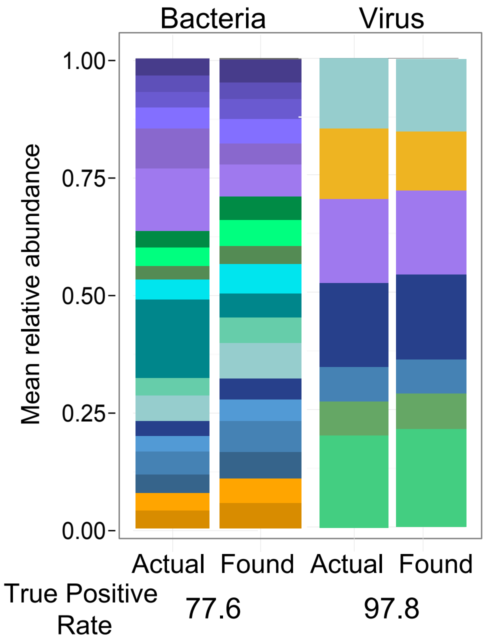


Figure 2.4. Clinical PathoScope validation with simulated communities.

Results from complex synthetic communities indicated that the target centric pipeline has an average true positive rate of 77.6 for bacteria species and 97.8 for virus. Relative abundance plots of the actual proportions and those found with Clinical PathoScope for bacteria and virus are shown.

2.4.4 Optimal Clinical PathoScope pipeline

The optimized Clinical PathoScope pipeline uses three reference genome libraries, four alignments modules and the original PathoScope read reassignment algorithm to identify pathogens in a given sample [Figure 2.5]. First, all reads from a sample are mapped against the reference genomes of the organisms of interest (*target library*, e.g. viruses) using up to the first 100 bases of each read. This initial alignment results in the removal of the greatest number of sequences by eliminating reads without strong sequence similarity to the target genomes. Second, reads that aligned to the target library are aligned against the reference library of the host species (*host library*) using the first 50 bases of each read. This step allows for any residual host contamination to be identified and removed from the set of candidate reads originating from the target genomes. Third, reads which did not align to the host library are aligned against additional reference genomes (*non-target library*) known to be negative targets of the analysis and which may overlap with the candidate read set. Similar to step one, reads are aligned using the first 100 bases of each read to maintain high specificity. Reads which did not align to the non-target library are realigned to the target library allowing up to *k* alignments (e.g., we recommend *k*=10 for viral detection) per read and subsequently passed to the read reassignment module in which reads with ambiguous alignments are reassigned to their putative correct genome of origin. In summary, any sequencing read contributing to the identification of a pathogenic genome must 1) align to the target genome library, 2) remain unaligned to the host genome library, 3) remain unaligned to the non-target library, and 4) retain its alignment to the target library. Finally, the pipeline produces a report detailing the number and proportion of reads originating from each genome identified in a given sample.

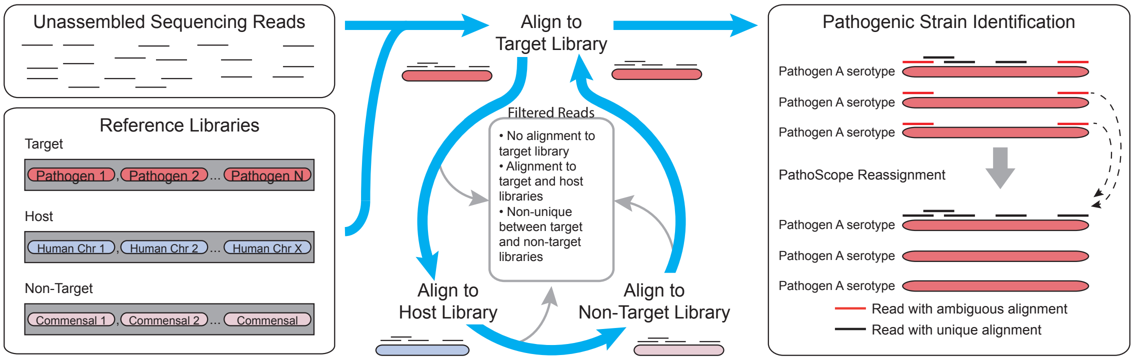

Unassembled Sequencing Reads
Reference Libraries
Target
Pathogen 1, Pathogen 2 ... Pathogen N
Host
Human Chr 1, Human Chr 2 ... Human Chr X
Non-Target
Commensal 1, Commensal 2 ... Commensal
Align to Target Library
Filtered Reads
• No alignment to target library
• Alignment to target and host libraries
• Non-unique between target and non-target libraries
Align to Host Library
Align to Non-Target Library
Pathogenic Strain Identification
Pathogen A serotype
Pathogen A serotype
Pathogen A serotype
PathoScope Reassignment
Pathogen A serotype
Pathogen A serotype
Pathogen A serotype
Read with ambiguous alignment
Read with unique alignment

Figure 2.5. Clinical PathoScope pipeline.

A computational subtraction method using varying sequence read lengths and ambiguous read reassignment. Unassembled sequencing reads are aligned against a target library containing reference sequences of the intended target(s) of identification (e.g. viruses). Reads aligned to the target library are then aligned to a host library. Any reads aligned to the host sequences are removed from further analysis. Next, reads are aligned against a library of known non-target sequences. Unaligned reads are then mapped back to the target library, allowing up to *k* alignments per read (e.g. k=10). These alignments are subsequently passed to an expectation maximization algorithm in which ambiguous alignments are reassigned to their most probable genome of origin. Upon reassignment, a report detailing the pathogens identified and their relative abundances is produced.

2.4.5 Software implementation and distribution

The Clinical PathoScope pipeline has been implemented in open-source Python, and is freely available for download at: http://sourceforge.net/projects/pathoscope/. The software requires the user to supply a fastq read file (after conducting quality control), any number of target, host, and non-target library Bowtie2 indices. Furthermore, the user has the option of changing the pipeline alignment parameters using inputs in the configuration file. For convenience, our viral, bacterial, and human alignment indices are freely available for download on the software distribution website. Clinical PathoScope will output two alignment files in SAM format, one directly from the Bowtie2 alignment, and another after read reassignment. Finally, the pipeline will output a tab-delimited summary report containing the genomes found in the sample as well as read numbers and proportions assigned to each genome.

2.4.6 Evaluation of Clinical PathoScope on clinical data

Four clinical datasets were utilized to evaluate the efficacy of Clinical PathoScope across a variety of scenarios [Table 2.1]. In addition, Clinical PathoScope was evaluated side by side with two previously published pathogen identification methods, RINS and READSCAN, on the basis of computational speed and accuracy at identifying pathogens in clinical sequencing samples.

		Average Run Time (minutes)		
Dataset	Target	Clinical PathoScope	RINS	READSCAN
Simulation	Virus	4.5	84.1	193.58
Simulation	Bacteria	13.1	1108.2	
PCCL	Virus	6.0	89.1	52.8
TMAdv	Virus	4.4	144.0	78.6
Mummy	Bacteria	25.0	1099	882

Table 2.3. Runtime comparisons of Clinical PathoScope and existing technologies.

Prostate Cancer Cell Line (PCCL)

Clinical PathoScope was able to rapidly decode the viral composition of this dataset; identifying the Human papillomavirus type 18 in fewer than 10 minutes. RINS and READSCAN both produced similar results; however, they required approximately four times the computational time to identify the pathogen, with run times of 89 minutes and 53 minutes, respectively [Table 2.3].

New World Titi Monkey Adenovirus Outbreak (TMAdv)

We examined Clinical PathoScope's performance in two clinical scenarios using the TMAdv dataset. First, to evaluate our pipeline in cases where the exact strain is missing from the target library, we excluded the TMAdv strain from the target library. In this scenario, Clinical PathoScope assigned reads to several adenovirus species **[**Figure 2.6A]. According to Chen *et al.*, the Simian adenovirus 3, which was the top ranked virus in the Clinical PathoScope result, is the closest phylogenetic relative to the TMAdv, with approximately 56% sequence similarity. Despite its highly divergent nature, Clinical PathoScope was able to successfully identify the closest phylogenetic neighbor of this novel species. Next, as a positive control, we included the TMAdv genome in our target library and validated that Clinical PathoScope accurately distinguished the TMAdv from all other adenovirus genomes [Figure 2.6B], identifying 12,568 reads from TMAdv. In their original analysis, Chen *et al.* used BLASTn (Altschul et al., 1990) to identify 16,524 reads from TMAdv. This discrepancy can be explained by the fact that BLASTn is a much more sensitive algorithm than Bowtie2. This moderate increase in sensitivity, however, results in a dramatic increase in run time, with BLASTn requiring ten times longer to complete the alignment than Bowtie2 when TMAdv is the only sequence in the database. Therefore, with rapid pathogen detection as the goal, a Bowtie2-based approach clearly provides a reasonable trade-off between speed and sensitivity, whereas if genome assembly is the goal, a BLAST-based approach might be preferable (at the cost of computational efficiency). Despite aligning approximately 4,000 fewer reads than the analysis in the original publication, we were still able to obtain 22.0x coverage of the TMAdv genome. While it is clear that Clinical PathoScope aligned substantially more reads with the TMAdv genome in the target library than in its absence, we were still capable of generating a list of candidate relatives with read counts proportional to their sequence similarity with the TMAdv. Furthermore, Clinical PathoScope completed analysis of this dataset in less than 5 minutes [Table 2.3].

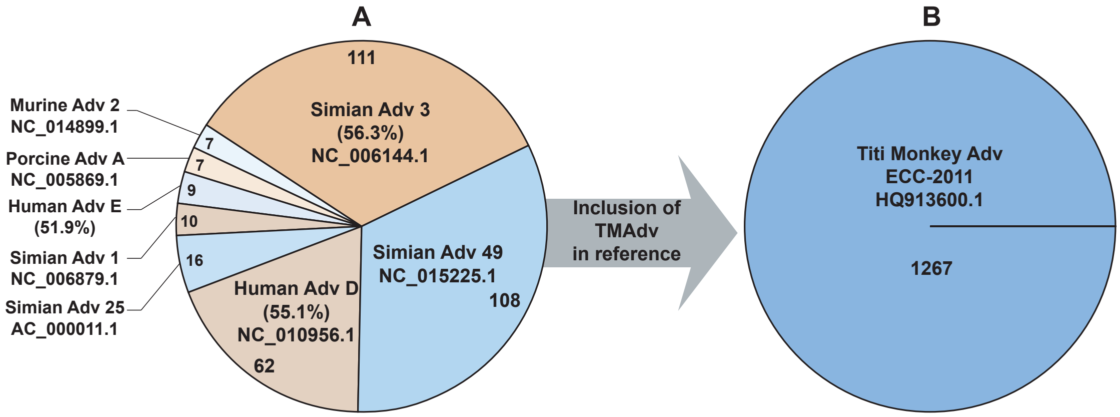


Figure 2.6. Alignment variations with and without TMAdv in the target library.

A) Without the TMAdv present in the target library, Clinical PathoScope assigned reads to several adenovirus genomes. The identified genomes are displayed according to the proportion of total reads aligned to all adenovirus genomes. The pairwise nucleotide identities of several adenovirus subtypes to the TMAdv genome according to Chen *et al.* are given in parentheses. The Simian adenovirus 3 had the most reads aligned of all adenoviral genomes, which is concurrent with its sequence similarity to the TMAdv. Additionally, the Human adenovirus D aligned the most reads of all human adenoviruses, which is concurrent with the analysis of Chen *et al.* B) Inclusion of the Titi Monkey Adenovirus (TMAdv) in the target library resulted in the assignment of 12,568 reads to the TMAdv reference genome.

With the TMAdv genome in the reference library, both RINS and READSCAN were able to accurately identify the correct viral genome in the sample. When the TMAdv was removed from the library, RINS generated a single contiguous sequence consisting of only 156 reads which mapped to 6 different adenovirus genomes, none of which included the nearest phylogenetic neighbor. This shows that, while assembly may be possible in a given sample, the ambiguous mapping of a contig to multiple genomes provides little information pertaining to the true subspecies of origin. Additionally, RINS required 144 minutes to complete its analysis of this dataset. READSCAN assembled several contigs of varying lengths and read counts from 16-60 reads per contig. However, the adenovirus strains identified and ranked by READSCAN based on their relative genome abundance score (Naeem et al., 2013) were inconsistent with phylogenetic relationships found by Clinical PathoScope and the original study (Chen et al., 2011b). Finally, READSCAN required approximately 80 minutes to analyze this dataset.

Tuberculosis in a Mummy

To demonstrate the performance of Clinical PathoScope with respect to bacterial pathogen identification, we analyzed a sample isolated from a mummy infected with tuberculosis. Using assembled contigs and comparative genomics, Chan *et al*. found evidence the deceased was infected with two *Mycobacterium tuberculosis* genotypes. Using patterns of deletions and SNPs, they concluded that both strains most closely resemble strain 7199/99, but also share similarities with strain H37Rv. When strain 7199/99 was included in the target database, Clinical PathoScope associates 32% of the reads with strain 7199/99 and 25% of reads with H37Rv. The majority of remaining reads were split between additional *M. tuberculosis* strains and *Nocardia* species. Chan *et al*. also identified *Nocardia* species using their assembly approach. Clinical PathoScope successfully identified the most closely related strains and furthermore, only required 25 minutes to complete the analysis. While these results are in agreement with the author's nearest-neighbor findings, we note that the number of *M. tuberculosis* strains in the sample (two unique strains according to Chan *et al.*) cannot be inferred from the Clinical PathoScope output alone. To successfully conclude the presence of two unique strains in the sample, a more complex, assembly based approach is required. Neither RINS nor READSCAN performed well on this dataset, requiring 1099.0 and 882.25 minutes, respectively, to complete the analysis, likely due to the large average read size of 297 bases and the complexity of the bacterial database. RINS assembled 20,483 unique contigs of varying length and reported 1,044,193 unique alignments of these contigs to 2,293 bacterial genomes. While vast, these results are uninformative as to the specific strains present within the clinical sample. Several contigs were assigned to various *M. tuberculosis* strains in the RINS report, however there was a tremendous lack of specificity with regard to the specific strains present in the sample. With thousands of other bacterial genomes identified and no metric for quantifying sequence abundance, the user is forced to interpret the results of thousands of contigs and millions of potential alignments, many of which are redundant or uninformative. READSCAN required less time to complete its analysis of the mummy dataset than RINS; however it also failed to generate a report detailing any of the identified pathogens. In their original publication, the authors demonstrate READSCAN primarily in the context of viral pathogen identification and note its performance improvements over previous methods. As can be observed from its run time on the mummy dataset, however, READSCAN has trouble scaling to larger bacterial datasets with many closely related strains of the same species.

Bacterial species identification from 16S amplimer aequencing

Clinical PathoScope was also tested on eight 16S amplimer samples (Accession: SRP028704), five originating from ATCC bacterial species, and three from patient tissue extracted from intensive care patients with suspicion of bacterial infections. As shown in Table 2.4, Clinical PathoScope was able to successfully identify the unique bacterial species in each of the first four ATCC samples with high accuracy. Furthermore, Clinical PathoScope was able to accurately identify the correct mixture of ATCC species in the fifth sample, assigning 30.4%, 30.2%, 21.2%, and 15.9% of the reads to *Escherichia coli*, *Enterococcus faecalis*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*, respectively.

For the three patient samples, we observed that the first sample (F1) contained a mixture of *Acinetobacter baumannii* (57.6%) and *Pseudomonas aeruginosa* (40.4%), and that the other two samples (G1 and H1) were dominated by *Pseudomonas aeruginosa* (94.6%) and *Enterobacter aerogenes* (84.2%), respectively. To validate these results, we constructed a phylogenetic tree of 16S genes from all genomes in the reference library that reside within the three genera identified in the clinical samples [Figure 2.7]. We then visually inspected the read coverage pileup plots of 16S genes that are unique between the species we identified and their phylogenetic neighbors [Figure 2.8]. We observed that read coverage is uniform across the genomes identified by Clinical PathoScope in each sample, resulting from the fact that they share 100% sequence similarity of their 16S genes. In contrast, we noticed large coverage gaps in the nearest phylogenetic neighbors, indicating that there were sequence variants in these regions that prohibited reads from aligning to these specific locations. This analysis further demonstrates the highly specific and accurate framework employed by Clinical PathoScope and its utility not only for strain-specific pathogen identifications, but also for 16S bacterial classification.

		Clinical PathoScope Results	
Accession	Sample Type	Species Identified	Reads Assigned (%)
SRR949994	*S. aureus* ATCC No. 25923 MSSA	*S. aureus*	3,479 (98.0)
		P. aeruginosa	36 (1.0)
SRR949995	*E. faecalis* ATCC No. 51299	*E. faecalis*	2,351 (89.8)
		S. aureus	139 (5.3)
		E. hirae	44 (1.7)
		P. aeruginosa	42 (1.6)
SRR949996	*P. aeruginosa* ATCC No. 27853	*P. aeruginosa*	5,661(82.3)
		E. coli	1,021 (14.9)
SRR949997	*E. coli* ATCC No. 25922	*E. coli*	4,169 (94.7)
		S. enterica	66 (1.6)
SRR949998	Mixture of *E. coli, E. faecalis,* *P. aeruginosa, S. aureus* (above)	*E. coli*	14,280 (31.9)
		E. faecalis	14,306 (31.9)
		P. aeruginosa	8,771 (19.6)
		S. aureus	6,594 (14.8)
SRR950015	Clinical Sample (F1)	*A. baumannii*	4,889 (59.4)
		P. aeruginosa	3,177 (38.7)
SRR950024	Clinical Sample (G1)	*P. aeruginosa*	1,131 (94.5)
		E. coli	45 (3.8)
SRR950025	Clinical Sample (H1)	*E. aerogenes*	587 (85.9)
		P. aeruginosa	18 (2.6)
		Erwinia sp. Ejp617	19 (2.8)
		E. coli	18 (2.6)
		S. enterica	9 (1.3)
		E. asburiae	10 (1.5)
		S. intermedius	8 (1.2)

Table 2.4. Clinical PathoScope performance on the 16S amplimer dataset.

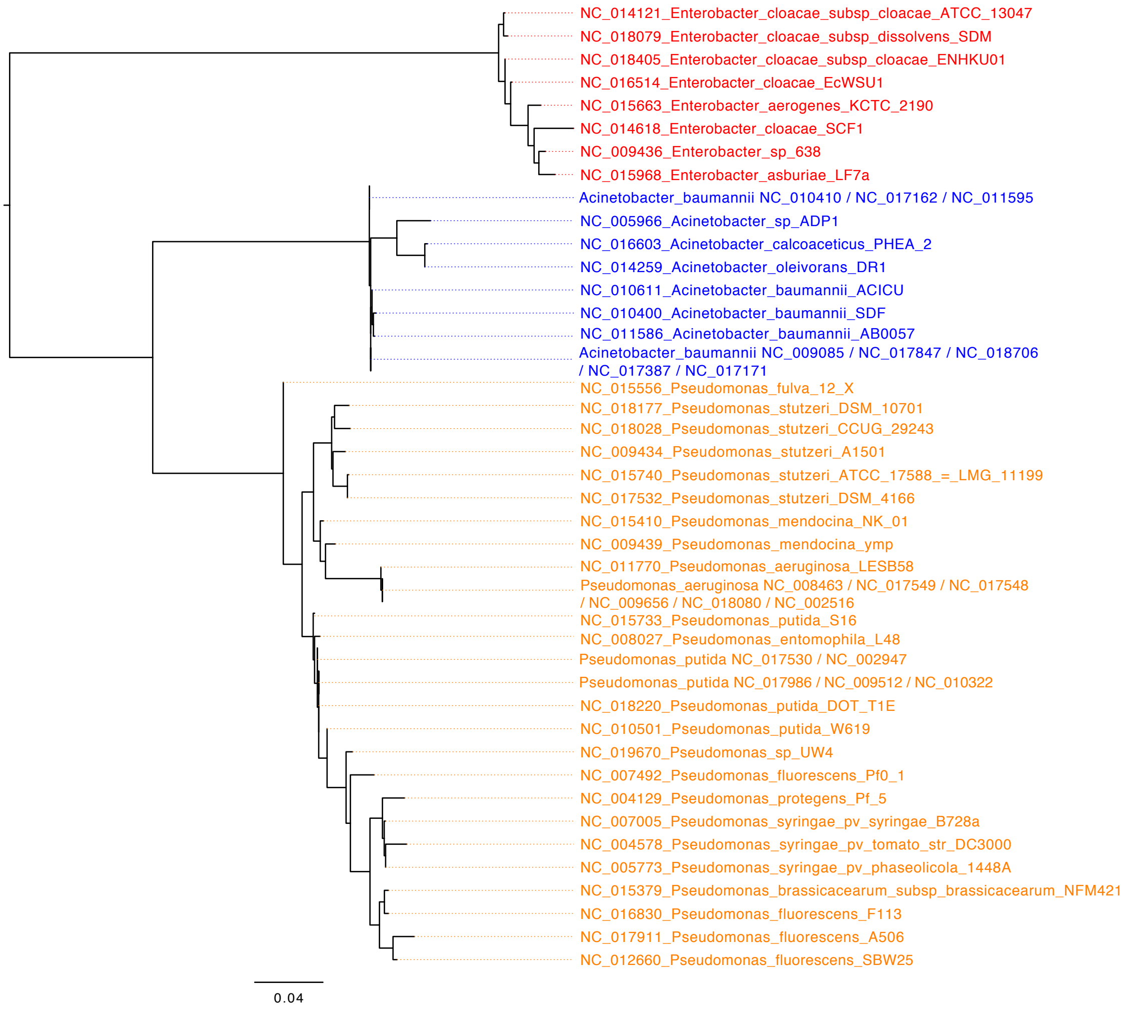
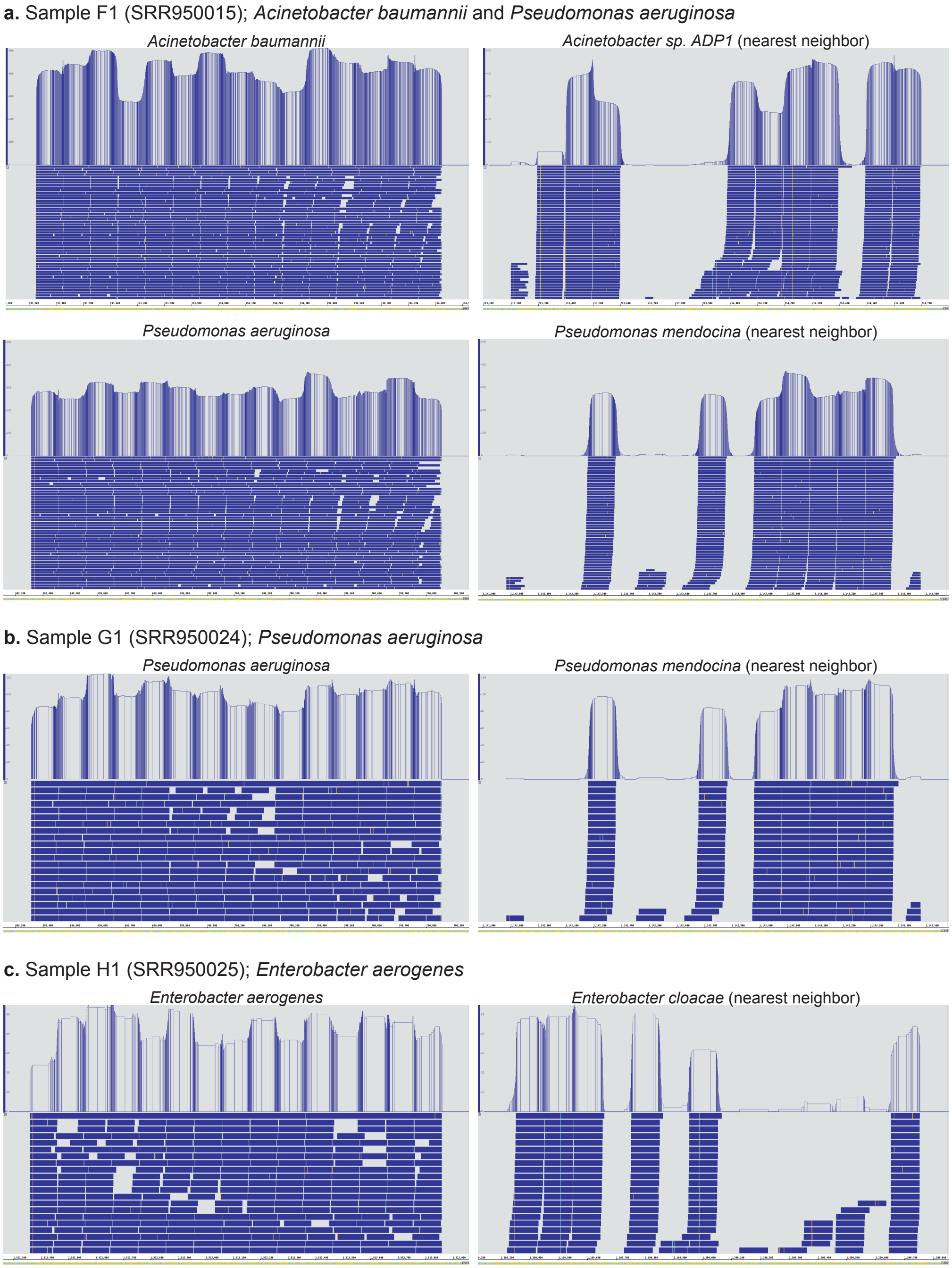

NC_014121_Enterobacter_cloacae_subsp_cloacae_ATCC_13047
NC_018079_Enterobacter_cloacae_subsp_dissolvens_SDM
NC_018405_Enterobacter_cloacae_subsp_cloacae_ENHKU01
NC_016514_Enterobacter_cloacae_EcWSU1
NC_015663_Enterobacter_aerogenes_KCTC_2190
NC_014618_Enterobacter_cloacae_SCF1
NC_009436_Enterobacter_sp_638
NC_015968_Enterobacter_asburiae_LF7a
Acinetobacter_baumannii NC_010410 / NC_017162 / NC_011595
NC_005966_Acinetobacter_sp_ADP1
NC_016603_Acinetobacter_calcoaceticus_PHEA_2
NC_014259_Acinetobacter_oleivorans_DR1
NC_010611_Acinetobacter_baumannii_ACICU
NC_010400_Acinetobacter_baumannii_SDF
NC_011586_Acinetobacter_baumannii_AB0057
Acinetobacter_baumannii NC_009085 / NC_017847 / NC_018706 / NC_017387 / NC_017171
NC_015556_Pseudomonas_fulva_12_X
NC_018177_Pseudomonas_stutzeri_DSM_10701
NC_018028_Pseudomonas_stutzeri_CCUG_29243
NC_009434_Pseudomonas_stutzeri_A1501
NC_015740_Pseudomonas_stutzeri_ATCC_17588_=_LMG_11199
NC_017532_Pseudomonas_stutzeri_DSM_4166
NC_015410_Pseudomonas_mendocina_NK_01
NC_009439_Pseudomonas_mendocina_ymp
NC_011770_Pseudomonas_aeruginosa_LESB58
Pseudomonas_aeruginosa NC_008463 / NC_017549 / NC_017548 / NC_009656 / NC_018080 / NC_002516
NC_015733_Pseudomonas_putida_S16
NC_008027_Pseudomonas_entomophila_L48
Pseudomonas_putida NC_017530 / NC_002947
Pseudomonas_putida NC_017986 / NC_009512 / NC_010322
NC_018220_Pseudomonas_putida_DOT_T1E
NC_010501_Pseudomonas_putida_W619
NC_019670_Pseudomonas_sp_UW4
NC_007492_Pseudomonas_fluorescens_Pf0_1
NC_004129_Pseudomonas_protegens_Pf_5
NC_007005_Pseudomonas_syringae_pv_syringae_B728a
NC_004578_Pseudomonas_syringae_pv_tomato_str_DC3000
NC_005773_Pseudomonas_syringae_pv_phaseolicola_1448A
NC_015379_Pseudomonas_brassicacearum_subsp_brassicacearum_NFM421
NC_016830_Pseudomonas_fluorescens_F113
NC_017911_Pseudomonas_fluorescens_A506
NC_012660_Pseudomonas_fluorescens_SBW25
0.04

Figure 2.7. Phylogeny of 16S genes for genera found in clinical samples.

We constructed a phylogenetic tree of 16S genes from all species in the reference library from the genera identified in the patient samples from the clinic. This tree was used to identify the nearest 16S neighbor of the Clinical PathoScope diagnosis, and to check initial mapping read coverage of 16S genes.



a. Sample F1 (SRR950015); *Acinetobacter baumannii* and *Pseudomonas aeruginosa*

b. Sample G1 (SRR950024); *Pseudomonas aeruginosa*

c. Sample H1 (SRR950025); *Enterobacter aerogenes*

Figure 2.8. Read coverage for 16S genes and nearest phylogenetic neighbors.

A) F1, B) G1, and C) H116S clinical samples (top frame: overall coverage, bottom frame: ‘pileup’ plot for a selected sets of the reads). Coverage for the ‘nearest’ phylogenetic neighbor contains large coverage gaps and some of the locations have mismatching bases for all reads. Combined these figures indicate that Clinical PathoScope has correctly identified the correct species in these clinical samples.

2.4.7 Comparison to metagenomic pipelines

Clinical PathoScope has been designed to facilitate a rapid and streamlined approach to identify strain-specific pathogens in noisy clinical sequencing samples. We compared our method directly with two previously published algorithms, RINS and READSCAN, which were designed specifically for pathogen identification in clinical samples. Additional methods, such as PathSeq (Kostic et al., 2011) and IMSA (Dimon et al., 2013), were also considered. These methods rely on several BLAT and BLAST alignments in order to filter sequencing reads which can take several hours to days to complete depending on the number of reads in a given sample. To evaluate these types of approaches, we implemented a similar BLAST-based workflow and applied this workflow to our second set of simulated samples with the bacterial library as the target. This approach resulted in a substantial decrease in performance with only 48.3% and 34.8% sensitivity for species and strain-specific identifications, respectively. This BLAST-based approach required 55 hours and 26 minutes, which is 300 times slower than Clinical PathoScope. Therefore, these algorithms are not practical methods for rapid clinical diagnostics.

We further expanded our comparisons to metagenomic pipelines that were not specifically designed for the identification of pathogens in clinical samples but whose methods or modules may be useful for the task. We first considered the CloVR-Metagenomics pipeline which clusters raw sequencing reads to reduce redundancy followed by a simultaneous BLASTX and BLASTN analysis against RefSeq and COG in order to annotate each sequencing read. CLoVR-Metagenomics does not address the issue of host contamination and thus wastes computational time clustering and annotating sequences originating from the host which can account for >90% of the clinical sample. While very sensitive, BLASTN is notoriously slow and does not scale well to large metagenomic samples (Ames et al., 2013), making CLoVR-Metagenomics impractical for rapid strain identification. Furthermore, the redundancy reduction procedures employed by CLoVR-Metagenomics collapse sequences with 99% nucleotide similarity which could potentially remove reads which distinguish two closely related strains of the same species.

We also considered assembly-based metAMOS (Treangen et al., 2013) and phylogeny-based LMAT (Ames et al., 2013). metAMOS offers a rich suite of assembly algorithms and pathogen annotation methods, however it does not incorporate any methods to remove host or contaminating sequences. As a result, the assembly of sequencing reads from a host-dominated clinical sample would require an attempt to assemble the entire host genome. This will result in a substantial and unnecessary increase in computational time and these contaminating reads could result in high instances of false positive mappings. LMAT, a software package designed for taxonomy classification, does not report strain-level annotation of sequencing reads nor does it report genome abundance information and thus cannot replicate the detailed pathogen report produced by Clinical PathoScope.

2.5 Conclusions

Sequence-based diagnostic tools have the potential to revolutionize the treatment of patients in the clinic, particularly those suffering from viral and bacterial infections. As the run times and error rates of modern sequencing technologies rapidly decline, it is essential that software be developed to analyze these data in a manner that is both fast and highly sensitive in order to provide physicians with the most accurate information possible. We have implemented a novel pipeline for pathogen identification that overcomes many of the challenges faced by current sequence-based methods including clinically appropriate run time and subspecies specific assignment of sequencing reads. We have also demonstrated our method's ability to identify multiple pathogens in a single clinical sample or the nearest phylogenetic neighbor of highly mutated or divergent species. Furthermore, Clinical PathoScope remained robust when analyzing datasets with lower than 1x coverage of the target genomes. It should be noted, however, that as coverage drops below 1x, the probability of sequencing a strain-specific segment of the target genome decreases. If these uniquely identifying reads are not sequenced and thus not present in the sample, Clinical PathoScope will tend to report the strain with the most aligned reads. Given that strain-specific reads do exist within a given sample, we expect the lower limit of coverage required to make a strain-specific identification to be comparable to our previously published results (Francis et al., 2013) in which we demonstrated the efficacy of our read reassignment algorithm with as low at 20% coverage of the genome

The reference genome libraries used in this analysis contain all sequenced and assembled viral and bacterial genomes from NCBI's RefSeq database. By avoiding genome assembly in lieu of more rapid computation, Clinical PathoScope is limited in that it can only identify pathogens that are present in these reference libraries. While the libraries used in this study characterize the majority of known pathogens, they do not contain draft genomes. To broaden and extend the application of Clinical PathoScope in future studies, we allow the user to exchange, modify, or extend these libraries as more data becomes available.

By comparison with existing methods, we have demonstrated that our method is the fastest strain-level pathogen identification algorithm currently available in the literature. As the number of sequenced pathogens grows, the breadth of the reference libraries used with Clinical PathoScope will increase, thus expanding the search space required to assign sequencing reads to a specific genome of origin. While this increase in search space will result in a linear increase in run time, we assert that our method will not lose its computational advantage over existing methods.

In addition to faster run times and more accurate results, Clinical PathoScope offers a user-friendly implementation. With only two dependencies, Bowtie2 and the PathoScope reassignment algorithm, Clinical PathoScope can easily be installed and run on a standard desktop computer, facilitating a simplified workflow for the accurate identification of pathogens in clinical sequencing samples. While designed for use by computational biologists and biologists, the reports produced by Clinical PathoScope may prove useful to physicians as they provide a complete picture of the microbial community of a given clinical sample which may influence clinical diagnoses and treatment options.

CHAPTER 3 Strain tracking bacteria in metagenomic samples of the skin microbiome

3.1 Abstract

Metagenomics, or the genomic sequencing of an entire community of microbiota (bacteria, fungi, virus), enables an investigation of the full complement of genetic material, including virulence, antibiotic resistance, and strain differentiating markers. The granularity to distinguish between closely related strains is particularly important as within one species some strains are beneficial while others are pathogenic to the host. A novel pipeline was developed to identify individual strains from complex metagenomic datasets. In this method, reads are first mapped against a database of all sequenced strains of an organism; the resulting alignment file is then processed by a Bayesian statistical framework which reports what percent of each strain is present in the sample. Finally, as further validation, the alignment file is parsed to identify reads mapping to informative single-nucleotide variants (SNVs). To validate the accuracy of the method, metagenomic samples were simulated composed of 6 and 12 *Propionibacterium acnes* strains, for which 78 sequenced strains are available. Results of the simulation study indicate the pipeline can successfully identify multiple different strains present in a sample with 96% sensitivity. When a strain was left out of the reference database, as is likely the case in real metagenomic samples, the pipeline correctly reported the most closely related strains the majority of the time. When the pipeline was applied to skin metagenomic samples from healthy volunteers, we found that individual strains are shared across multiple body sites (e.g.; arm crease, forearm, nose,…) of the same person. These strain tracking tools

provide the framework for future investigations comparing longitudinal datasets or investigating strains specific to a disease state.

Note: The majority of work presented in this chapter has been previously published in (Oh, Byrd et al., *Nature* 2014).

3.2 Introduction

Skin is the first defense against pathogenic bacteria while simultaneously harboring billions of commensal bacteria. These symbiotic skin bacteria play important roles in lipid metabolism, inhibiting colonization by transit bacteria, education of the immune system, and pathogen suppression. They inhabit defined topographical regions of the skin, such as the arm pit, elbow crease, forehead, toe webs, and heel (Grice et al., 2009). A clinician can easily and discretely access these individual sites for sampling. This highlights an advantage of the skin for microbiome studies over the gut that has the same topography, but is easily assessed only with the aggregate stool sample. Thus intra-individual comparisons are possible with skin samples that are not possible with stool. The ability to compare different sites is particularly valuable for the study of common skin disorders, which show predilection for stereotypical skin sites such as eczema inside the elbow versus psoriasis on the outside of the elbow(Kong et al., 2012) (Paulino et al., 2006). There is evidence that these common skin disorders and others including acne and rosacea are associated with altered microbial states (Fitz-Gibbon et al., 2013). Before studying disease states, it is important to establish a baseline for healthy individuals.

To date the majority of skin microbiome studies explore microbial composition based on amplicon sequencing of universal marker genes, such as the16S ribosomal RNA (rRNA) gene for bacteria and the Intertranscribed Spacer (ITS) of rRNA for fungi (Findley et al., 2013; Grice et al., 2009). Based on the 16S rRNA survey, it was found that bacterial colonization is dependent on the physiology of the skin site with specific bacteria being associated with the moist, dry, and sebaceous microenvironments. Sebaceous sites are dominated by lipophilic *Propionibacterium* species, while humidity loving *Staphylococcus* and *Corynebacterium* species are most abundant in moist areas. In contrast to colonization patterns found for bacteria, results of ITS surveys indicate fungal diversity is more dependent on body location than physiology. Fungi of the genus *Malassezia* dominate core-body and arm sites, while foot sites are colonized by a more diverse combination of *Malassezia*, *Aspergillus*, *Cryptococcus*, *Rhodotorula*, *Epicoccum* and others [Figure 1.3].

To expand upon previous amplicon surveys, clinical and laboratory techniques were developed to sample healthy volunteers and extract enough biomass for shotgun metagenomics sequencing. Metagenomics, or genomic sequencing of an entire community of microbiota (bacteria, fungi, virus), enables an investigation of the full complement of genetic material, including virulence, antibiotic resistance, and strain differentiating markers. Under IRB protocol (08-HG-0059; PI: Segre), samples for complete microbial and human genome sequencing were collected from 18 body sites of 15 healthy adult volunteers, 6 females and 9 males with ages ranging from 23 to 39. Samples were sequenced on an Illumina Hi-Seq to yield 30-100 million 100 basepair paired-end reads. In total, 244 samples were available for analysis.

3.3 Methods and Results

3.3.1 Multikingdom metagenomics

With no well-validated pipelines or databases, analysis of skin metagenomic samples was dependent on the curation of a whole genome database with known skin microbes and the development of new computational tools. To address these needs, first a comprehensive multi-kingdom database was compiled from 2,342 bacteria, 389 fungi, 1,375 virus, and 67 archeael genomes sequences from the National Center for Biological Information (NCBI), the Human Microbiome Project (HMP), the Saccharomyces Genome Database (SGD), the Fungal Genome Initiative (FGI), and FungiDB (Stajich et al., 2012). Where multiple genomes for a reference were available, the complete genome was selected for inclusion over draft genomic sequences. Next to analyze the metagenomic data, Clinical Pathoscope (Byrd et al., 2014), a pipeline to rapidly and accurately remove host contamination, isolate microbial reads, and identify potential disease-causing pathogens, was modified to simultaneously detect bacteria, fungi, archaea, and viruses [Figure 3.1]. Modifications to the original pipeline include removing the initial target-mapping step, removing the non-host filtration steps, and the addition of a genome coverage calculation. Given there were 4 large target databases, the initial target-mapping step was less effective in reducing runtime than when there were 1 or 2 target databases, thus it was excluded.

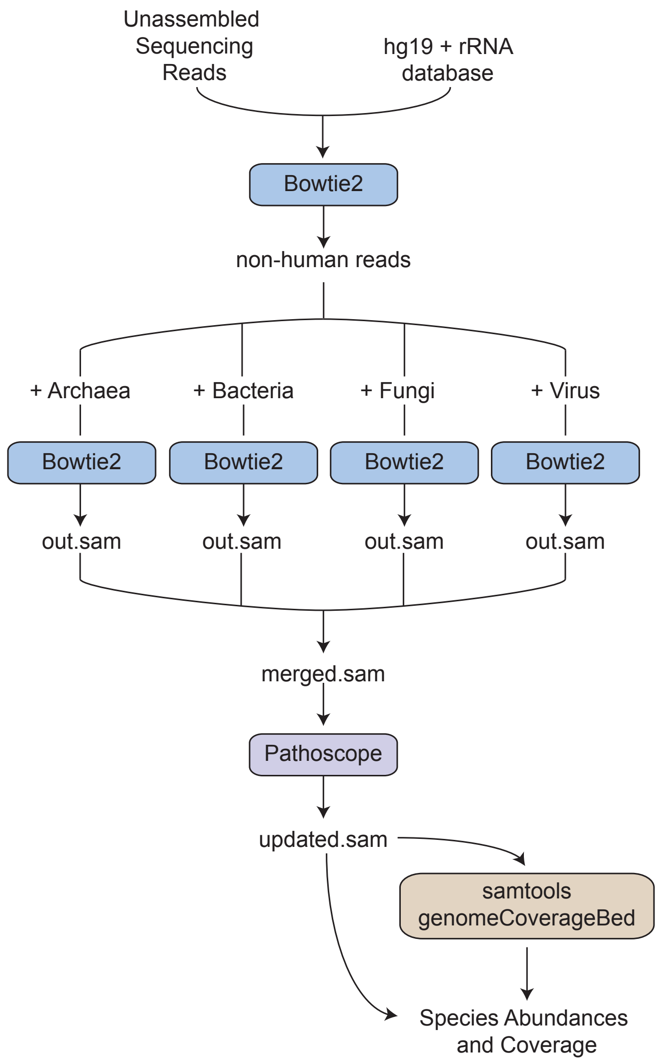


Figure 3.1. Complete metagenomic analysis pipeline.

To analyze a sequencing sample, first, reads mapping to the human database are filtered away. The remaining non-human reads are then individually mapped with Bowtie2 to archaeal, bacterial, fungal, and viral databases. The resulting sam alignment files are then combined and processed with the Pathoscope read assignment algorithm. The updated alignment file is then processed with samtools to determine the percent of the genome covered for each species in the sample.

In the updated pipeline, reads were first mapped to the human hg19 reference genome and human rRNA sequence using bowtie2's -very-sensitive parameter. Reads mapping to human were not considered further in this microbe centric analysis. Human-derived reads varied from 5-99% of the total reads depending on skin site and stochastic features [Figure 3.2]. Remaining non-human reads were then mapped to the microbial genome collection using bowtie2's (Langmead and Salzberg, 2012) -very-sensitive and –k 10 parameters such that the top 10 hits were retrieved. Multiply mapping reads were then reassigned using Pathoscope v1.0 (Francis et al., 2013), which uses a Bayesian framework to examine each read's sequence and mapping quality within the context of a global reassignment. Read hit counts were then normalized by genome length and scaled to sum to one. Coverages of each genome were calculated using the genomeCoverageBed tool in the Bedtools suite(Quinlan and Hall, 2010). For relative abundance and diversity calculations, genomes with coverage < 1 were removed. The remaining genomes' relative abundances were subsequently rescaled to one.

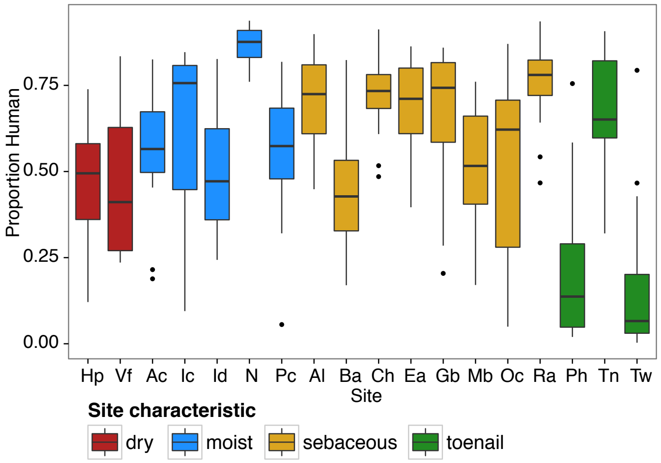


Figure 3.2. Percent human by body site.

Boxplots (line indicates median; boxes represent first and third quartiles) show, for each site, % reads mapping to human hg19 that are discarded before analysis. Sites are colored by site characteristic. Sites label as in **Figure 3.5**.

For validation, taxonomic assignments of bacteria and fungi were compared to 16S and ITS amplicon results, as well as to the output from a bacterial and archaeal mapping tool, Metaphlan (Segata et al., 2012). The Pathoscope pipeline's results were highly correlated with Metaphlan results. For species counts, the correlation was ρ = 0.96 [Figure 3.3A]. For genus level relative abundances, the correlation was ρ > .90 for all genera compared [Figure 3.3B]. The Pathoscope amplicon correlations were also high; ρ > 0.85 for all bacterial genera compared and ρ > 0.64 for all *Malassezia* species [Figure 3.3C,D].

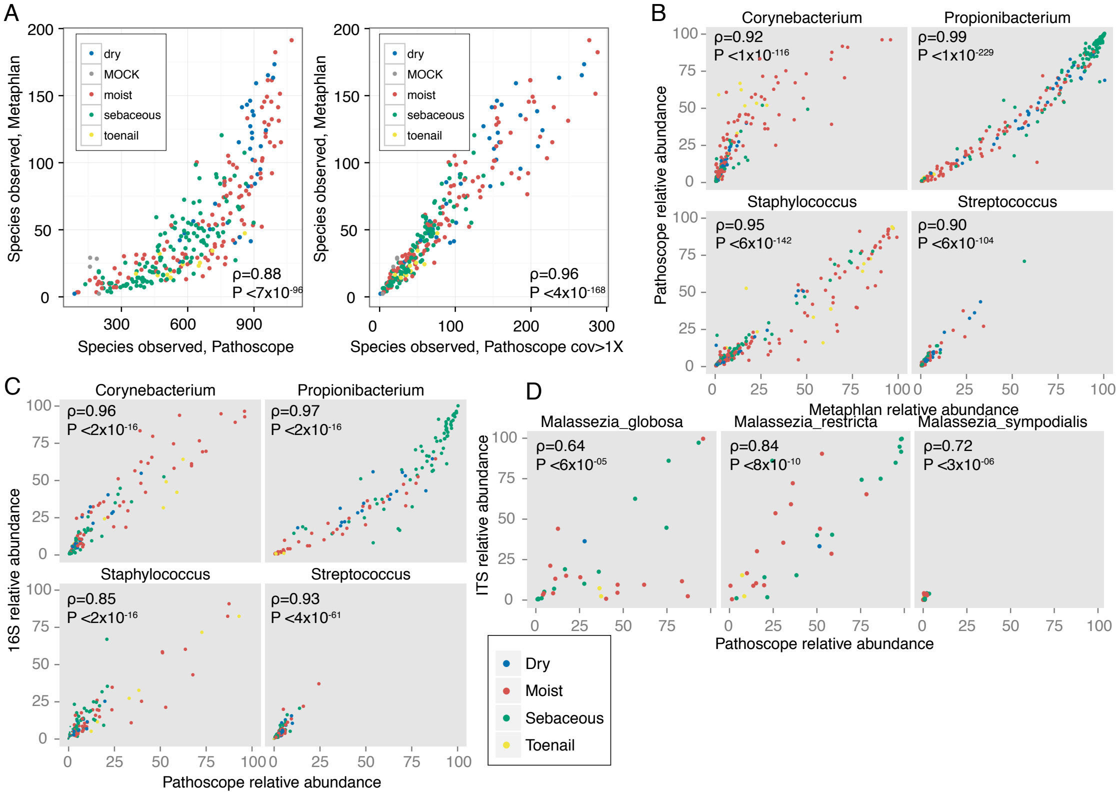


Figure 3.3. Comparison between Pathoscope and Metaphlan.

A) Number of species observed with no coverage cutoff (Left) and a coverage cutoff off of > 1 (Right). B) Relative abundance of the bacterial genera. C,D) Comparison between Pathoscope and amplicon results in regards to C) Relative abundance of bacterial genera with 16S. D) Relative abundance of fungal species with ITS**.**

When all healthy volunteer samples were run through the pipeline, bacteria and fungal abundances and distributions were similar to what was seen in the previously published amplicon surveys. The distribution of bacteria was mostly driven by the characteristic of the site, and the distribution of fungi was driven by the body location. *P. acnes* was the dominant bacteria in sebaceous regions, and *Malassezia* was the dominant fungi throughout the core body sites [Figure 3.4].

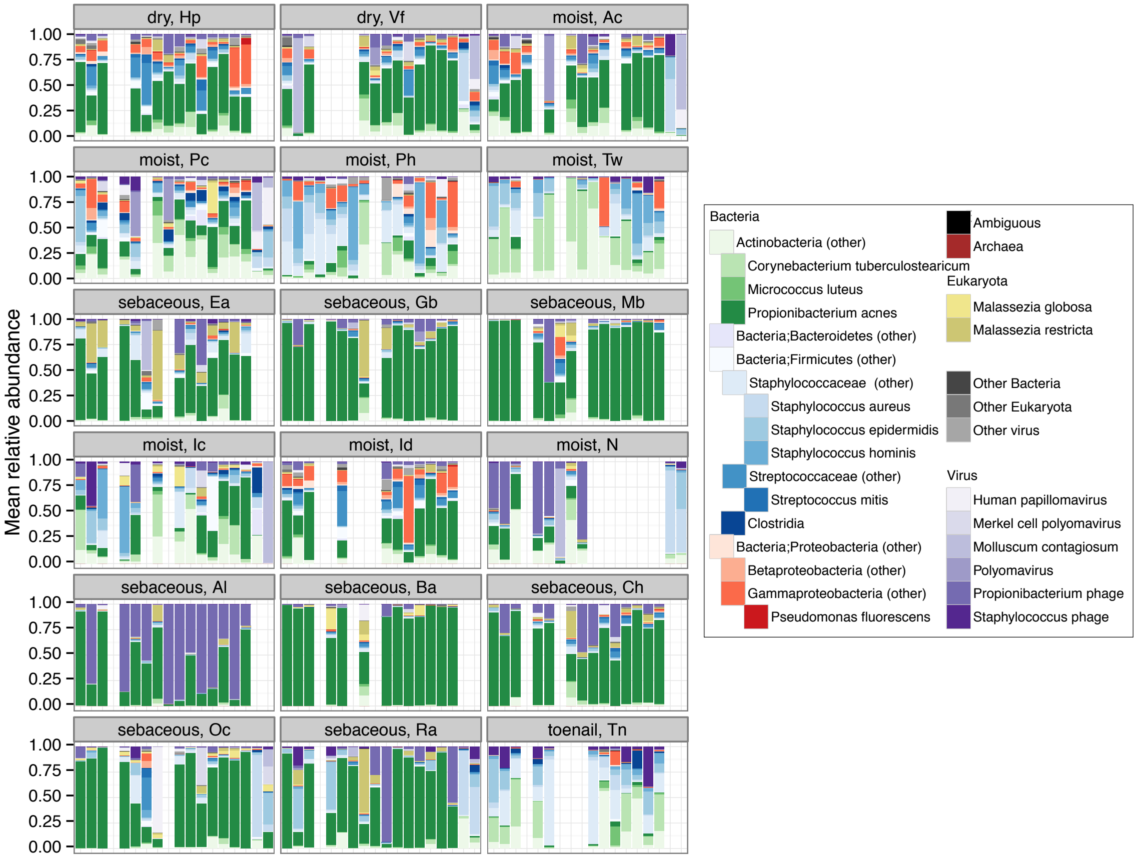


Figure 3.4. Multikingdom analysis across healthy individuals by site.

Relative abundance of prominent skin taxa in healthy volunteers. Within a site, each bar represents the microbes present in an individual.

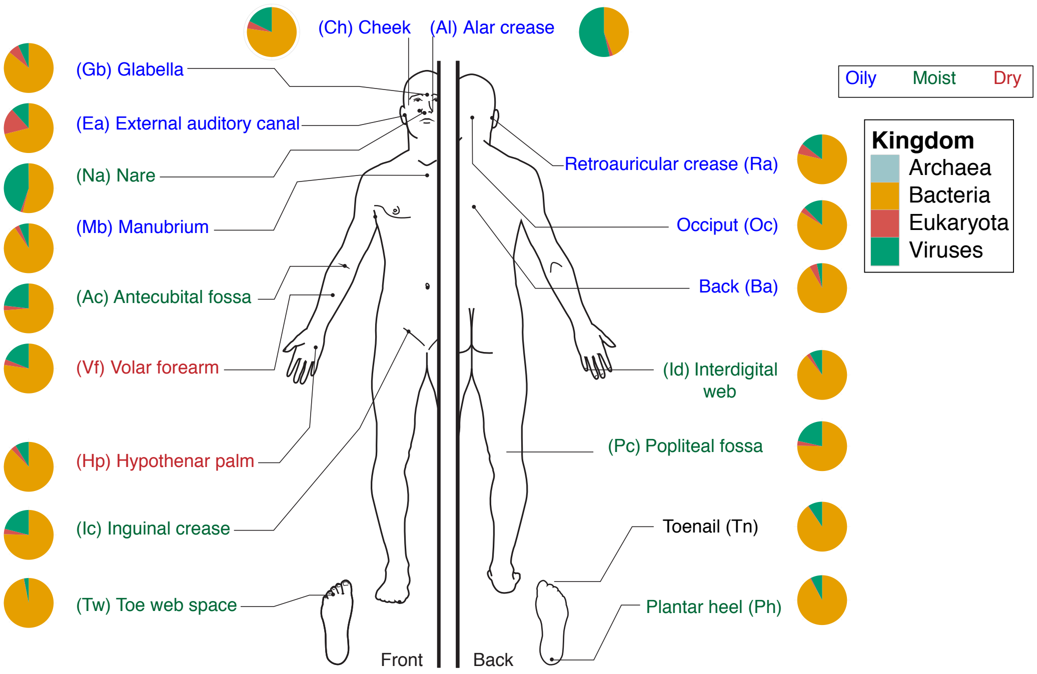


Figure 3.5. Relative kingdom abundances across different skin sites based on metagenomic survey.

Consensus pie charts show the relative abundances of archaea, bacteria, eukaryotes, and viruses across 18 different body sites. Body sites are colored based on their physiological classification.

New observations include relative abundances of the kingdoms [Figure 3.5]. The majority of samples were primarily bacteria, but a handful of samples, particularly the nares and alar crease, had a high proportion of viruses. Those viruses were most often bacteria phage, specifically Staphylococcus and Propionibacterium phage. The average fungal abundance at all sites was low, less than 10 percent; even on the feet where the fungal diversity was much higher than the rest of the body. Caveats to concluding low fungal abundance include the lack of high quality full fungal genomes, and the difficulty lysing fungi with a standard DNA extraction method.

3.3.2 Strain-level metagenomics

Marker based studies are limited in taxonomic resolution to genus or species level. Functional information can be inferred from marker genes (Langille et al., 2013), but strain differences are lost. The importance of differences between strains is highlighted by the increasing number of studies identifying large numbers of non-core genes within the pangenome of a species (Conlan et al., 2012; Tomida et al., 2013). While previous studies have used 16S rRNA gene profiling and SNPs to study intra-species variation between and within individuals (Schloissnig et al., 2013), comparisons of strains across multiple different body sites using reference sequences have not yet been done. To differentiate between closely related strains of the same species, parameters within the metagenomics pipeline were adjusted such that the stringency of bowtie2 was increased and Pathoscope's tendency to choose a parsimonious list was reduced [Figure 3.6A]. Strain tracking was focused on *P. acnes* and *S. epidermidis* species because of their high abundance in the skin and the availability of sequenced genomes.

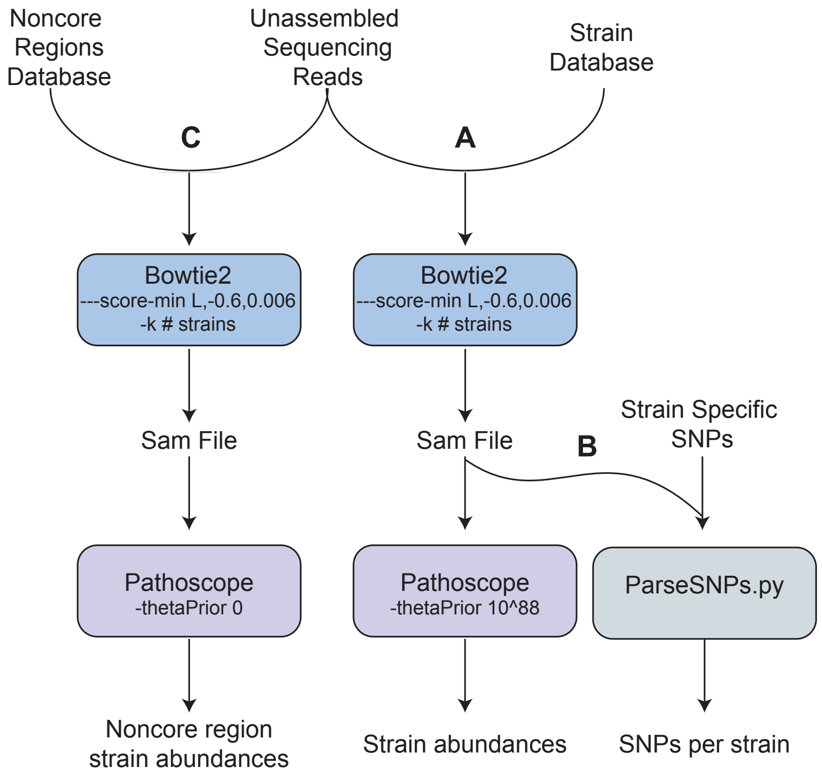


Figure 3.6. Three tested strain tracking approaches.

To detect strains present in metagenomic samples three alternative strategies were tested. A) whole genome + Pathoscope utilized a database composed of all sequences genomes of a species B) strain specific SNPs were identified and used to determine strain presence. C) noncore regions + Pathoscope utilized a database composed of only the noncore, variable, regions for each strain of a species.

The *P. acnes* and *S. epidermidis* databases were curated from all complete and draft genomes present for these species at NCBI, totaling 78 and 61, respectively. Isolates HL037PA2, HL037PA3 HL044PA1, and SK182B-JCVI were excluded from the *P. acnes* database as they likely represent different *Propionibacterium* species (Butler-Wu et al., 2011; McDowell et al., 2012). Reads were mapped to each species-specific database using bowtie2 with the most stringent parameters (--score-min L,-0.6,0.006), allowing zero mismatches and as many hits as genomes in the database (-k 78 or –k 61). This stringent criteria is necessary given 88% of the *P. acnes* genome is core, defined as regions shared between all reference genomes, and 80% of the *S. epidermidis* genome is core (Conlan et al., 2012; Tomida et al., 2013). Read assignment using Pathoscope was performed as described for the metagenomics pipeline, except theta_prior, an option that changes whether reads are assigned to as few genomes as possible, was set to 10^{88} (most genomes permitted). Read hit counts were then normalized by genome length and scaled to sum to one. Strains with a normalized abundance less than 1 percent were grouped as "Other".

To evaluate the ability of Pathoscope to accurately reassign reads to very similar strains, sensitivity was assessed using complex synthetic communities and it was demonstrated that the presence of unique genomic loci can allow discrimination between subtypes [Figure 3.6B,C]. To create a comprehensive test set, 3 sets of 6 different strains were chosen for both *P. acnes* and *S. epidermidis*. Strains were chosen such that all major taxonomic clades were represented. For each species, the 3 sets were then combined into all possible combinations to have samples of 12 and 18 genomes. For each of the sets, 50,000, 100,000, and 500,000 reads per genome were simulated with 5 different seeds using the simulator Mason (parameters: mason illumina -s ## -N ## -sq -n 100 -i -hs 0.0 -hi 0 -hnN –nN) (Holtgrewe, 2010); given those parameters, Mason incorporates an Illumina sequencing error rate into the reads. Thus 45 samples of 6 genomes, 45 samples of 12 genomes, and 15 samples of 18 genomes were tested for each species. In addition to the samples with equal abundances per strain, the 3 sets of 6 different strains were simulated to have staggered abundances ranging from 5,000 to 500,000 reads per genome.

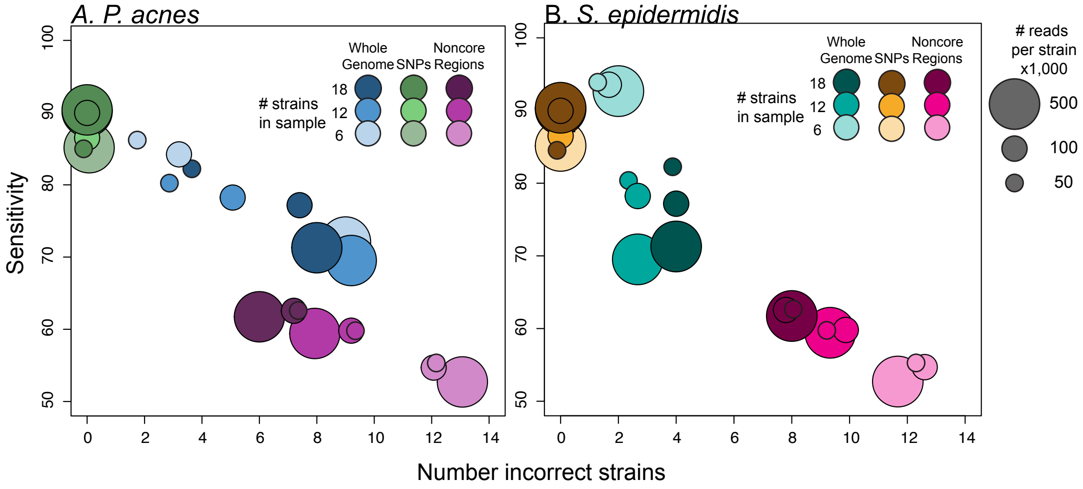


Figure 3.7. Strain-tracking simulation results.

A) *P. acnes* B) *S. epidermidis*. Size of the circle indicates the number of simulated reads per strain in the sample, the larger the circle the greater the number of reads. Different color families indicate which strain-tracking pipeline was used (whole genome, SNPs, or noncore regions). Shades of a color represent the number of different strains present in a sample, the darker the shade the more strains that were present.

Results of the simulation study indicate that the strain-tracking pipeline has an average sensitivity of ~78% for *P. acnes* and ~92% for *S. epidermidis* when considering abundance across all simulated samples [Figure 3.7]. This discrepancy in sensitivity exists because *S. epidermidis* genomes are less similar to each other than *P. acnes* genomes giving Pathoscope more regions of dissimilarity to accurately resolve strains. When only considering presence/absence of strains, the sensitivity is ~98% for *P.acnes* and ~99% for *S. epidermidis*. While the pipeline rarely misses strains present in the sample, it does falsely report strains as being present. The number of incorrect strains increases as the number of reads per genome increases. This is caused by the slow accumulation of sequencing errors in places that perfectly match incorrect strains. The majority of falsely reported strains are closely related phylogenetically to the strain identified. Using clade membership as a proxy for relatedness, we see the *P. acnes* sensitivity increase from ~78% for strain level to ~91% when using clade levels previously established by Tomida *et al.* (Tomida et al., 2013). Such specific clade identifications have not been previously published for all *S. epidermidis* strains.

To further validate the strain-tracking pipeline, two alternative pipelines were created using SNPs and noncore regions to discrimination between subtypes. SNPs unique to a strain in core regions of the genome were identified using nucmer (Delcher et al., 2002) and custom scripts. Nucmer was also used to identify non-core regions in each of the genomes. KPA171202 and SK137 were used as the *P. acnes* references, and ATCC_12288 and RP62A were used for *S. epidermidis*. To increase confidence, only SNPs and noncore regions identified with both references were used for strain identification. In agreement with previous studies (Conlan et al., 2012; Tomida et al., 2013), 88% of the *P. acnes* genome was identified as core and 80% of the *S. epidermidis* genome. To visualize relationships between the strains, all SNPs identified in core regions were used to create phylogenetic trees with the program PhyML (Guindon et al., 2009)[Figure 3.10B, Figure 3.11B]. In the SNP-based pipeline, custom scripts parsed the bowtie2 sam file, generated in the previously described strain-tracking pipeline, for reads mapping to SNPs unique to a strain [Figure 3.6C]. Counts per strain were then normalized based on the total number of SNPs per strain. When tested on the simulated reads, the average sensitivity of this approach is 88% for *P. acnes* and 93% for *S. epidermidis* with almost no strains falsely identified [Figure 3.8]. In the non-core region-based approach, the whole genomes databases in the Pathoscope-based pipeline were replaced with databases composed of only the non-core regions for each strain [Figure 3.6C]. Because the similarity between strains is reduced when using only noncore regions, the Pathoscope -thetaPrior variable was decreased to 0. Using only the non-core regions, the average strain-tracking sensitivity falls to 55% for *P. acnes* and 65% for *S. epidermidis,* and the average number of incorrect strains increases respectively to 11 and 8 [Figure 3.7]. This drop in sensitivity and increase in false positives indicate that differences in noncore regions alone are not sufficient for strain level analysis.

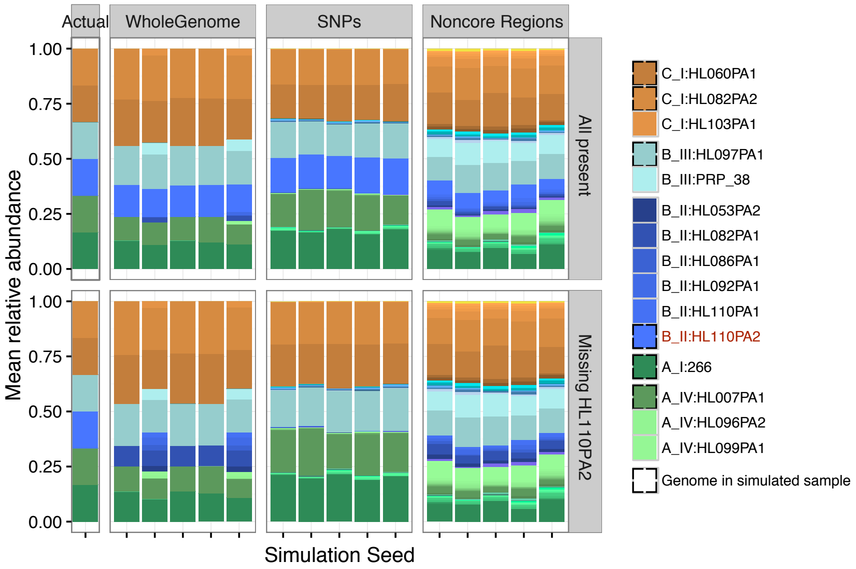


Figure 3.8. *P. acnes* simulation results for 3 strain tracking approaches.

Relative abundance plots show how accurate each of the 3 strain tracking approaches is when all sequenced strains are in the database (top) and when a strain (HL110PA2) is missing from the database, the more realistic scenario. Similarly colored bars represent closely related strains, or strains residing in the same phylogenetic clade. Strains present in the actual sample are boxed in black. For each approach, results are shown for 5 samples simulated with different seeds.

To evaluate the behavior of each pipeline when a strain is missing from the database, as is likely the case in real metagenomic samples, a set of *P. acnes* simulated samples were run through each pipeline with one of the strains removed from each database. Under these more realistic circumstances, the whole-genome approach identified closely related strains of the missing strain [Figure 3.8A]. Using clade membership as a proxy for relatedness, when the correct strain is present in the database the clade-level sensitivity is 87.9% percent. When the strain is missing, the clade-level sensitivity remains high at 86.6% percent. This validates the whole-genome approach's ability to successfully report the most closely related strains; however, because the reads are mapping to several closely related strains, the exact number of strains present cannot be accurately determined, only that something within that clade is present. Although the SNP-based approach performed better than the whole-genome based approach when all strains are present in the database, it is unable to identify closest neighbors when a strain is missing from the database [Figure 3.8B]. Because the SNP pipeline only uses SNPs that are unique to a strain, relatedness between strains cannot be inferred from SNPs alone. On the other hand, with the non-core region based approach it is possible to identify closest neighbors [Figure 3.8C]. By incorporating both SNP and noncore information into its reassignment method, the whole genome Pathoscope-based pipeline is able to exploit the high sensitivity of the SNP approach and the nearest-neighbor tracking ability of the non-core region based approach.

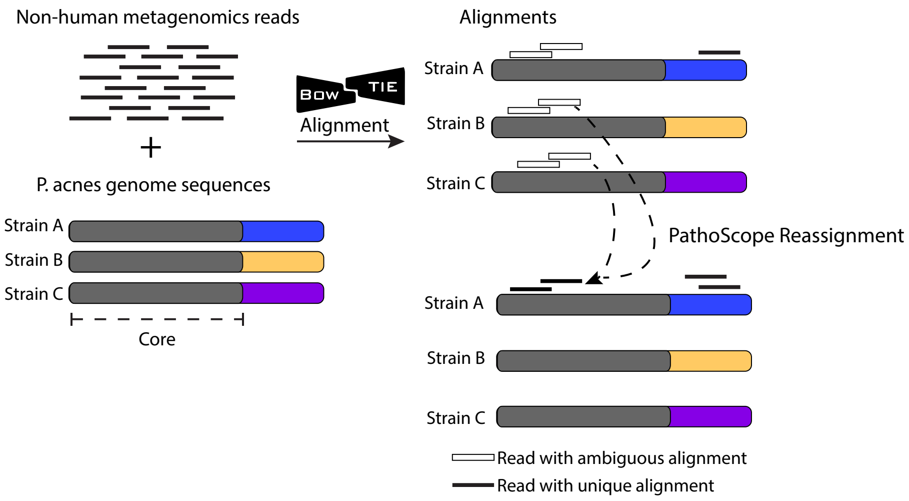


Figure 3.9. Strain tracking pipeline.

Nonhuman reads are mapped against a database of all sequenced strains of an organism. Multiple mapping reads are reassigned to the most likely genome of origin with the Bayesian statistical framework Pathoscope.

Based on positive results from the simulated data, all metagenomic samples were run through the whole-genome Pathoscope based pipeline [Figure 3.9]. At the majority of sites, humans were colonized by heterogeneous communities of *P. acnes* and *S. epidermidis* strains [Figure 3.10, Figure 3.11]. Similarity between samples was assessed using the Yue-Clayton theta similarity index that considers both presence/absence of strains and their relative abundance [Figure 3.12]. These results showed that an individual’s communities of *P. acnes* strains were more similar across his or her body sites than between individuals. Similar trends of intra-individual similarity were also observed across core body sites with *S. epidermidis*. Distinctly, for all individuals in this study, strains from clade B of *S. epidermidis* predominated the feet (orange strains in Figure 3.11). Interestingly, in a comparative analysis of *S. epidermidis* genomes (Conlan et al., 2012), this group was labeled as the professional commensals, as no nosocomial isolates, those isolated from hospital infections, clustered in this group. Overall, these results suggest that the host and the environment can differentially shape commensal strain communities. Further analyzes at this resolution will be powerful in tracking microbial communities across time in steady state and disease conditions.

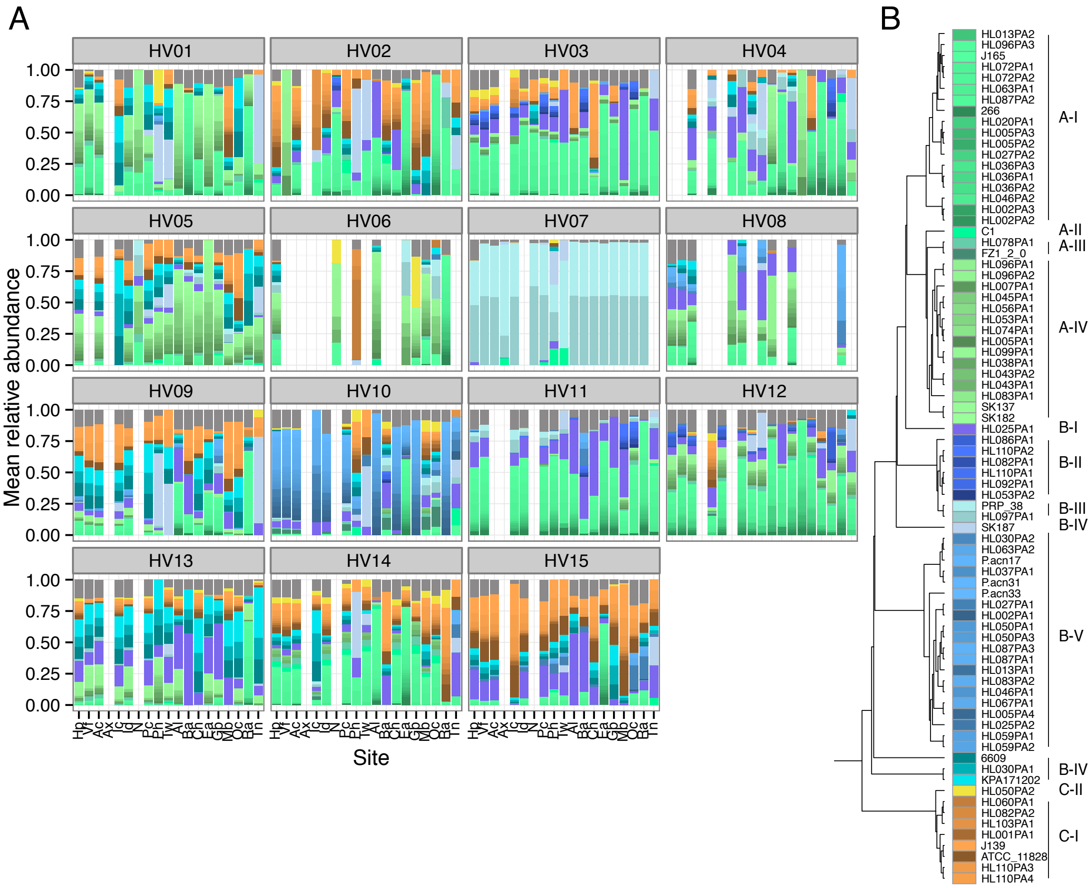


Figure 3.10. *P. acnes* strain tracking across body sites.

A) Full strain-level assignments for samples with relative abundances of closest related *Propionibacterium acnes* strains, by individual. B) Dendrograms of strain similarity. Trees were generated using core SNPs. Bar of colors indicates delineations of subtypes where phylogenetically more similar genomes are in similar colors; for example, we defined 12 subtypes for *P. acnes*.

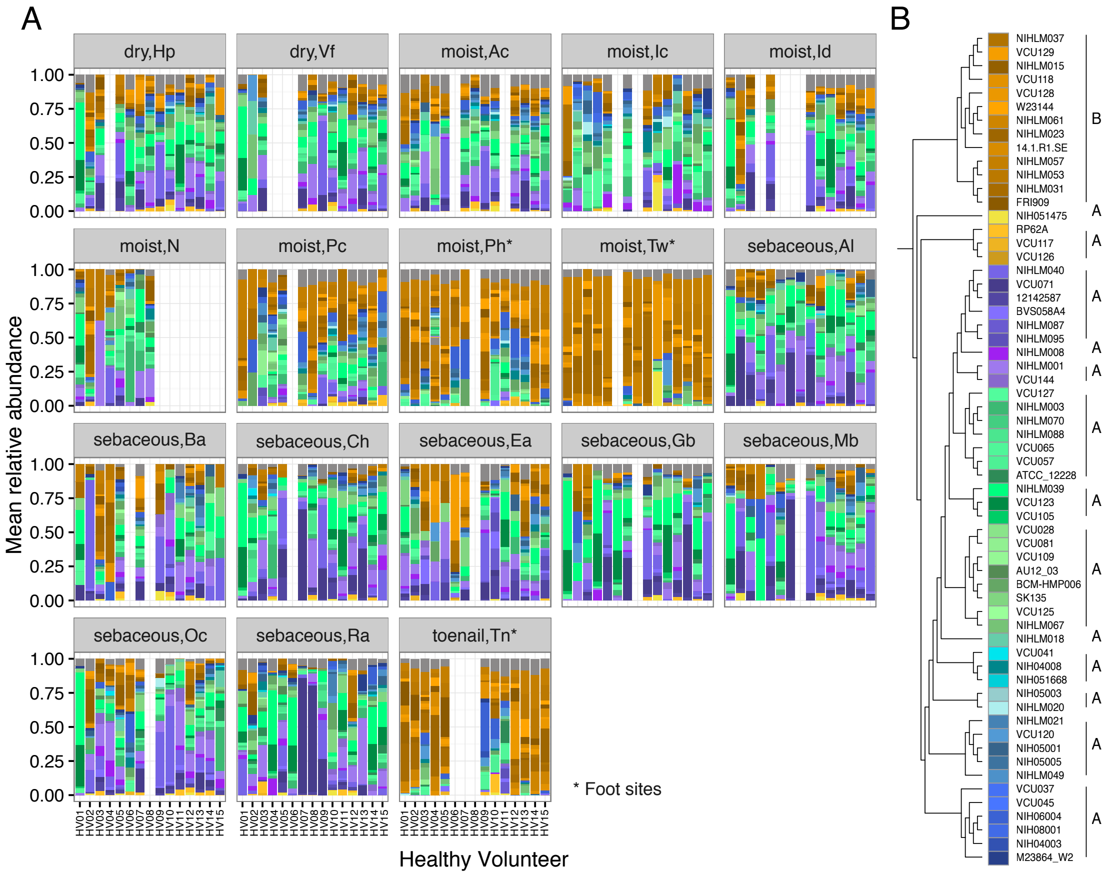


Figure 3.11. *S. epidermidis* strain tracking across body sites.

A) Full strain-level assignments for samples with relative abundances of closest related *Staphylococcus epidermidis* strains, by body site. B) Dendrograms of strain similarity. Trees were generated using core SNPs. Bar of colors indicates delineations of subtypes where phylogenetically more similar genomes are in similar colors; for example, we defined 14 subtypes for *P. acnes*.

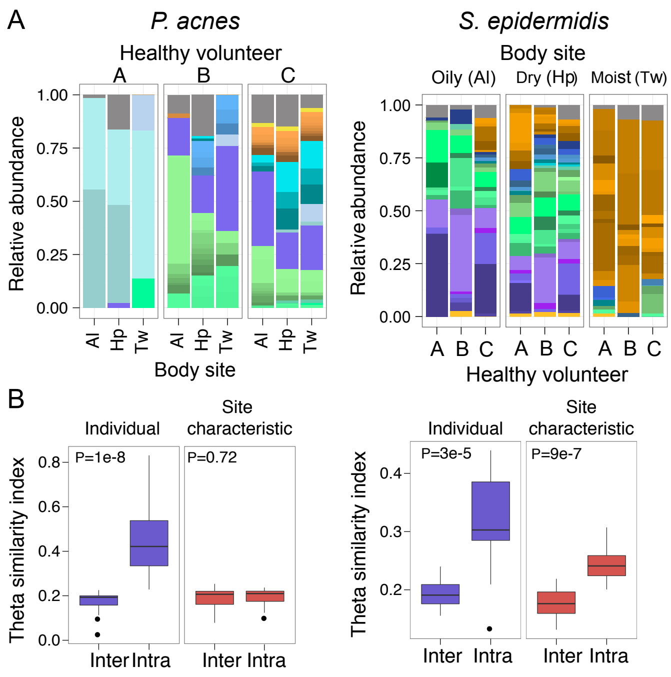


Figure 3.12. *P. acnes and S. epidermidis* are differentially shaped by host and the environment.

A) Relative abundance plots of *P. acnes* (left) and *S. epidermidis* (right) at representative sites. Similarly colored bars represent closely related strains as in [**Figure 3.10**B, **Figure 3.11**B]. *P. acnes* subtypes differ more significantly between individuals than site characteristic, while *S. epidermidis* subtypes differ by site characteristic and individual. B) The Yue-Clayton theta calculates similarity between two samples based on both the number of shared features and their relative abundances. θ=0: dissimilar; θ=1: identical.'Inter' reflects similarity between individuals or site characteristics; 'intra' reflects similarity to samples within individuals or site characteristic.

CHAPTER 4 Temporal stability of the human skin microbiome

4.1 Abstract

Biogeography and individuality shape the structural and functional composition of the human skin microbiome. To explore these factors' contribution to skin microbial community stability, we generated metagenomic sequence data from longitudinal samples collected over months and years. Analyzing these samples using a multi-kingdom, reference-based approach, we found that despite the skin's exposure to the external environment, its bacterial, fungal, and viral communities were largely stable over time. Strain and single nucleotide variant level analysis showed that individuals maintain, rather than reacquire prevalent microbes from the environment. Longitudinal stability of skin microbial communities generates hypotheses about colonization resistance and empowers clinical studies exploring alterations observed in disease states.

Note: The majority of work presented in this chapter has been previously published in (Oh[*], Byrd[*] et al., *Cell* 2016).

4.2 Introduction

Human skin is the first line of defense against pathogens while simultaneously harboring a diverse milieu of commensals including bacteria, fungi, and viruses. These symbiotic organisms play essential roles in lipid metabolism, colonization resistance to transient organisms, and education of the immune system (Belkaid and Segre, 2014; Grice, 2015; Scharschmidt and Fischbach, 2013). Previous studies have shown a strong site-specificity to microbial community composition and function: the physiologic characteristics of a skin site, including pH, temperature, moisture, sebum content, and topography shape the local microbial community (Costello et al., 2009; Findley et al., 2013; Grice et al., 2009; Grice and Segre, 2011; Oh et al., 2014). Understanding community variability across skin sites has provided the foundation to study the corresponding site-specificity to disease predilection, for example atopic dermatitis (eczema) in the bends of the arms and legs (Kong et al., 2012) and psoriasis on the elbows and knees (Alekseyenko et al., 2013). However, it is poorly understood why disease predilection changes over human lifespans and whether fluctuations in host-intrinsic factors, such as immunity or hygiene, influence microbial community composition and function. Understanding stability determinants is critical to studies investigating if homeostatic forces contribute to a healthy skin microbial community and if alterations influence host health.

In addition to skin's biogeography, as defined by physiologic factors such as sebaceous, moist, or dry, individual discriminatory attributes also likely contribute to skin microbial community dynamics over time. Using the high resolution and multi-kingdom analyses afforded by metagenomic shotgun sequencing, we have shown that low abundance microbial species including bacteria, fungi, and viruses can differentiate between individuals. Observing inter-kingdom dynamics is meaningful since these interactions may exacerbate disease severity (Peleg et al., 2010) or facilitate transitions from opportunistic to pathogenic. Moreover, a subspecies-level analysis of dominant skin species showed that strains can be unique to an individual, while the population heterogeneity of other species can be more specific to skin physiology (Oh et al., 2014).

Distinguishing between strains of the same species is necessary because some strains are beneficial while others may be pathogenic to the host.

Shotgun sequencing data provides resolution difficult to achieve with phylogenetic marker gene analyses. Metagenomic surveys of the gut have shown that single nucleotide variants (SNVs) (Schloissnig et al., 2013) and gene copy number variations (Greenblum et al., 2015; Zhu et al., 2015) can identify individual-specific strains, which illustrate functional differences that cannot be explained by species composition alone. Longitudinal studies in the gut have found individual-specific strains persist for a year (Schloissnig et al., 2013) or more (Faith et al., 2013). These studies have leveraged a combination of compositional and functional attributes to identify a broad trend of long-term retention of one's individual strains over time, even at the SNV level.

To understand community dynamics of healthy human skin, we investigated the temporal stability and diversity of skin microbial communities, expanding our previous metagenomic study to re-sample individuals at successive timepoints. Here, we show that community stability persists regardless of sampling time interval and despite constant exposure of skin communities to extrinsic factors [Figure 4.1]. Interestingly, the nature and degree of this stability is highly individual-specific; a trend previously observed with phylogenetic marker gene sequencing across body sites (Gajer et al., 2012). Strain-level analysis reveals that this stability is driven primarily by the maintenance of individual strains over time, rather than through the acquisition of prevalent microbes from the environment and other individuals (Lax et al., 2014). We present new insights into how biogeography and individuality regulate stability and transience of the skin microbiome community.

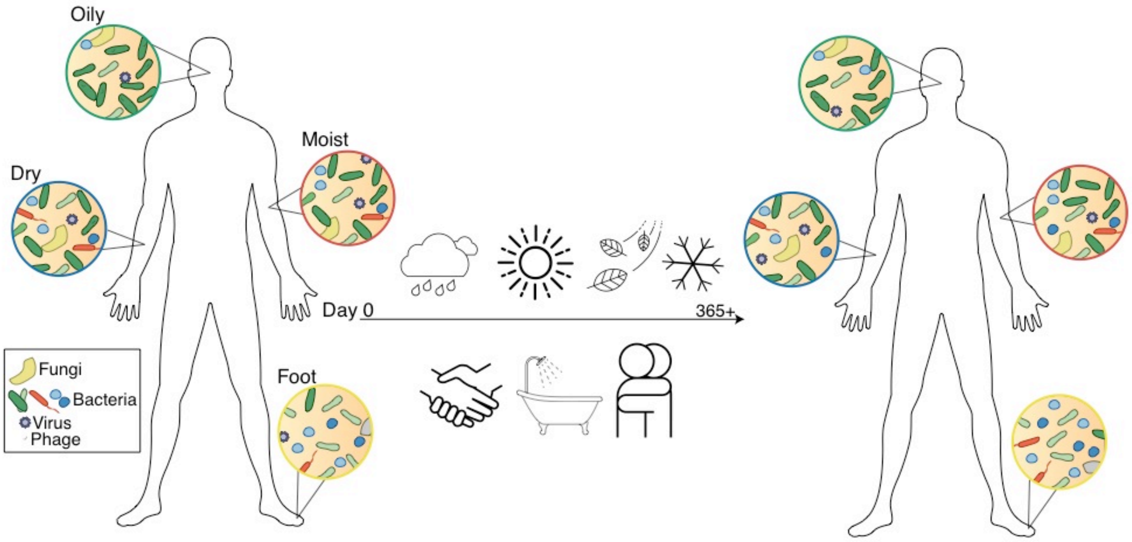


Figure 4.1. The skin microbiome is largely stable overtime despite environmental exposures.

4.3 Results

4.3.1 Skin microbes are largely stable at a community level

Our previous studies defined that the diversity and composition of skin microbial communities possess both site- and individual-specific qualities (Oh et al., 2014). To assess the effect of time on these characteristics, we collected samples over long (1-2 years) and short (1-2 months) time intervals. 12 healthy individuals were sampled across 17 skin sites at 3 timepoints for a total of 594 samples and 720 Gbp of shotgun microbial sequence data [Figure 4.2A]. For taxonomic reconstructions, we mapped microbial reads to a multi-kingdom reference database. To assess the stability of skin microbial communities, we compared community membership and structure over short and long time intervals using the Yue-Clayton theta index, which calculates the distance between communities based on relative proportions of shared and non-shared species in each population (Yue and Clayton 2005). $\theta = 0$ indicates dissimilar and $\theta = 1$ identical communities. We observed that an individual's short- and long-term community similarity significantly exceeded similarity between individuals [Figure 4.2B], similar to observations in gut and other communities (Caporaso et al., 2011; Costello et al., 2009; Faith et al., 2013; Flores et al., 2014; Human Microbiome Project, 2012). At all sites, long-term was lower than short-term similarity at the species level; a trend also observed when comparing timescales of 1 day versus 3 months (Costello et al., 2009). Bacterial and fungal communities of sebaceous sites were the most stable regardless of time intervals [Figure 4.2C]. Surprisingly, dry sites including high-exposure, high-perturbation sites like the palm, were also stable regardless of time. Foot sites were the least stable, with significant differences over both short- and long-term. This may be due to a combination of behavioral and physiologic factors, including shoe-wearing habits, personal hygiene, or features such as the thickness of plantar stratum corneum (the upper layer of skin).

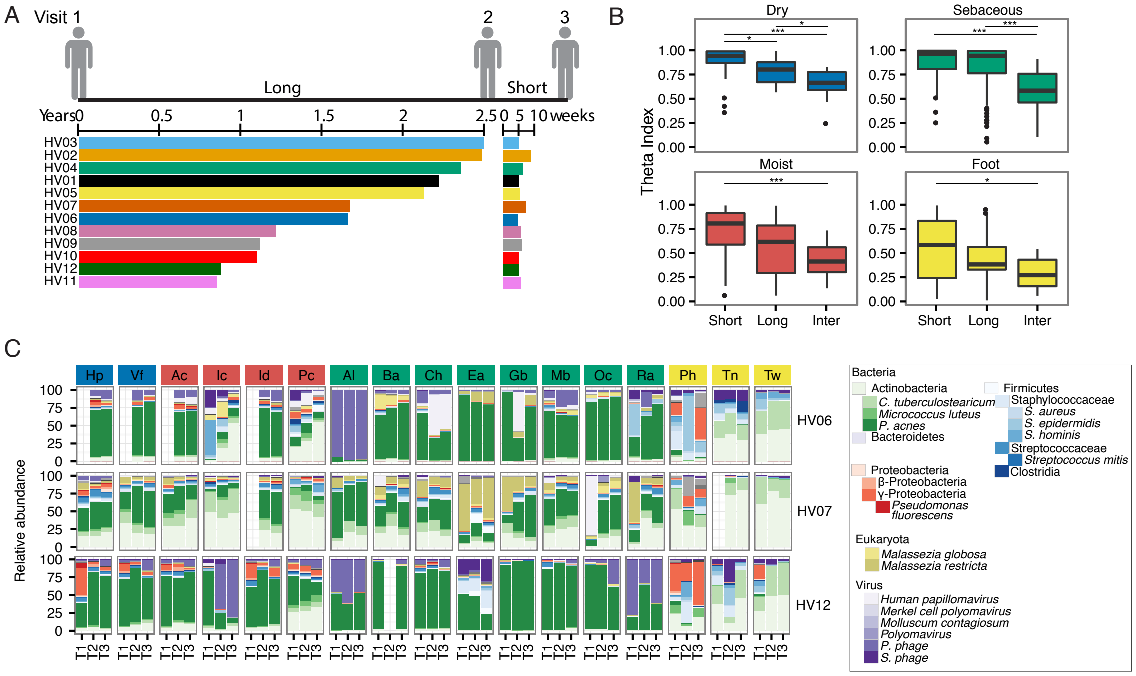


Figure 4.2. Study design and community stability between timepoints.

A) Longitudinal study design to show time between samplings. B) Boxplots of Yue–Clayton theta indices calculate similarity between sites aggregated by characteristic. ‘Long’ duration indicates ~1-2 years between samplings; ‘Short’ duration averages a month (T2 to T3). For comparison, ‘Interpersonal’ values show the average between individuals. Bonferroni-adjusted **P* <0 .05 value, ***P* <0 .01 value , ****P* < .001 value. C) Relative abundances of the most common skin bacteria, fungi, and viruses are shown for three representative individuals.

4.3.2 Individuals have distinct microbial SNV signatures that are stable over time

Shotgun metagenomic data empowers unprecedented resolution of microbial communities at the strain and SNV level. To explore stability at this high resolution, we focused on the prevalent skin bacterial commensals *Propionibacterium acnes* and *Staphylococcus epidermidis*, which have dozens of available genome sequences (Conlan et al., 2012; Tomida et al., 2013). Previously, by mapping shotgun metagenomic reads to this reference set of phylogenetically diverse genomes, we identified that skin sites harbor complex subspecies variation for both *P. acnes* and *S. epidermidis*. Strain heterogeneity could be unique to an individual (*P. acnes*) or specific to skin site (*S. epidermidis*) (Oh et al., 2014).

Here, we observed that *P. acnes* strains are remarkably stable over time across body sites [Figure 4.3A,B; Figure 4.4]. We quantified this stability with the Yue-Clayton theta similarity index [Figure 4.3C], taking into account both strain presence/absence and relative abundance. Temporal stability, short- or long-term, surpassed the similarity between individuals, indicating that *P. acnes* stability likely derives from the maintenance of an individual's strains over time and less from the acquisition of new strains from the environment or other individuals.

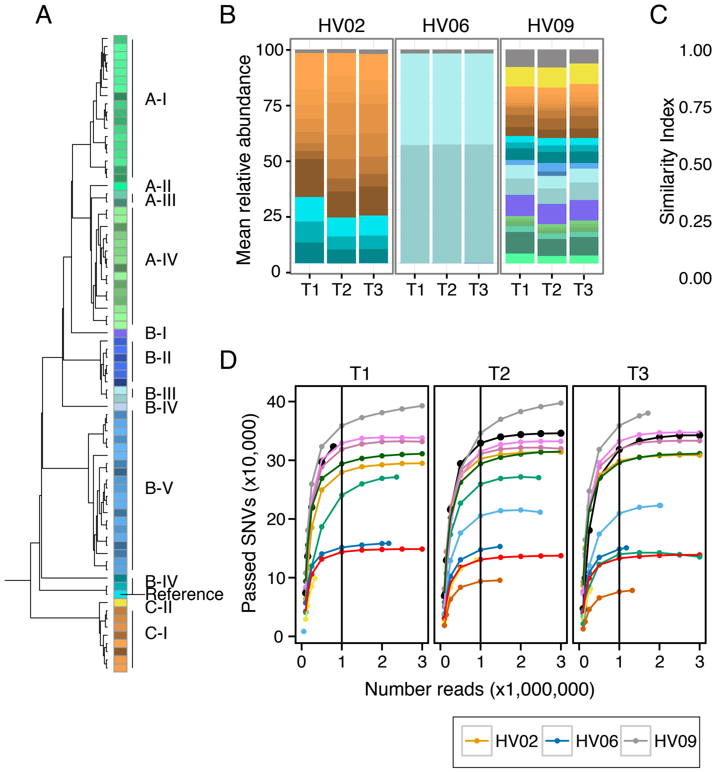


Figure 4.3. Individual-specific strain and SNV signatures are stable over time.

A) Dendrogram of *P. acnes* strain genome similarity based on core SNVs. B) *P. acnes* strain relative abundance plots of 3 representative individuals' manubrium, colors as in (A). Full set of strain classifications is shown in **Figure 4.4**. C) Boxplots of Yue-Clayton (left) and Jaccard (right) theta indices indicate similarity between strains (all body sites) and SNVs (manubrium and back) of *P. acnes* in a time series (θ = 1 is identical). ****P* < .001 value, Wilcoxon rank-sum test. D) Rarefaction curves demonstrate core SNV accumulation with read subsampling for manubrium sites. For remaining panels, SNVs are reported for samples subsampled to 1 million reads. Colors correspond to individuals as shown in **Figure 4.2**A.

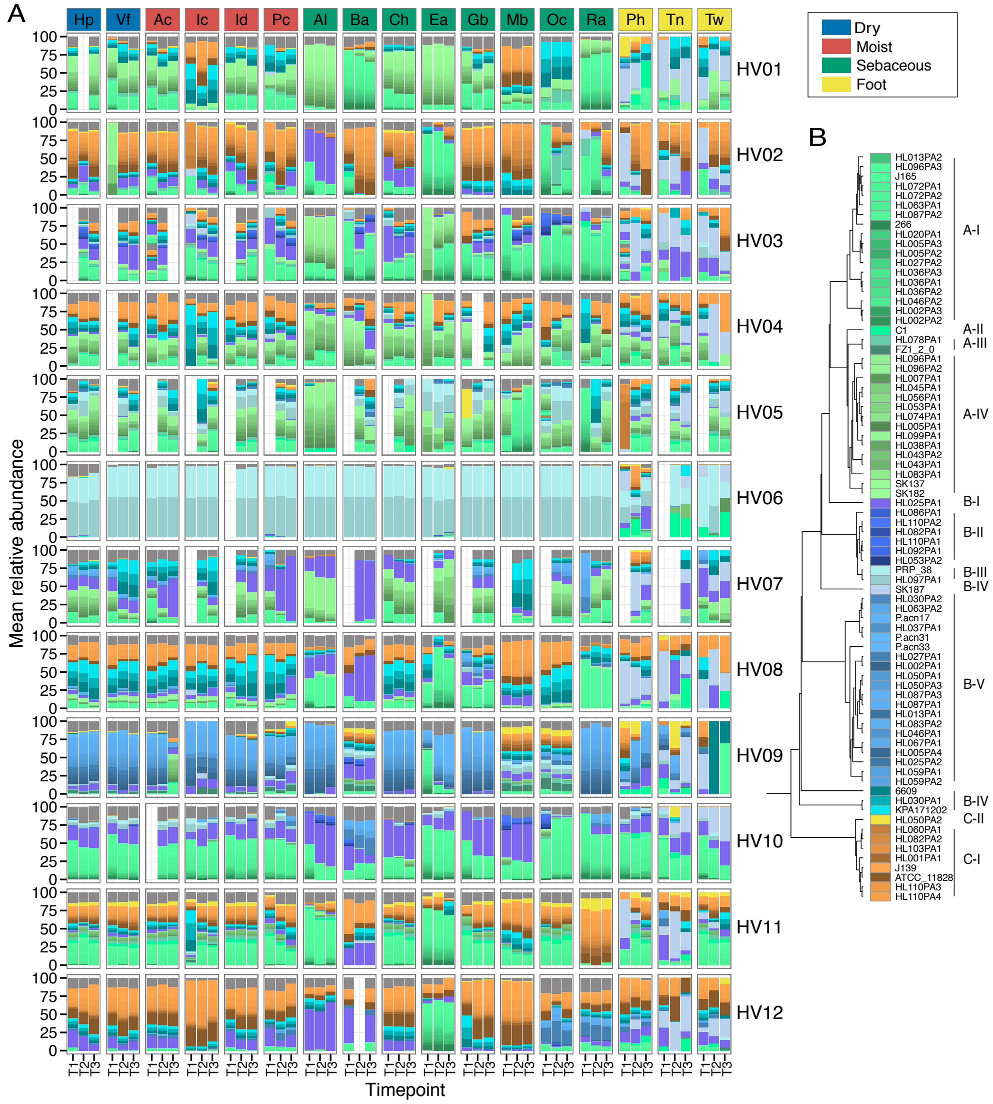


Figure 4.4. *P. acnes* strains are stable over time across body sites.

A) Full *P. acnes* strain tracking for all individuals, all sites. Sample headings are colored by site characteristic. T1, T2, and T3 indicate timepoints, with T1 and T2 being long-duration timepoints (>1 year) and T2 and T3 short-duration timepoints (~1 month). Colors correspond to those in (B). B) Dendrogram of *P. acnes* strain similarity based on core SNVs. Similar strains are grouped into clades.

To validate this hypothesis and to assess within-subject retention of strains, we also examined shared SNVs in the *P. acnes* genomes across the longitudinal samples. To power this analysis, we focused on two sebaceous sites, the manubrium and back, which have high sequencing depth and *P. acnes* abundance [Figure 4.5]. For these sites, we used SNVs specific to the *P. acnes* core genome (2,248,676 bps region of the genome shared between all 78 sequenced strains) to ensure even representation of SNVs in every metagenomic sample. We identified 83,081 variant positions in the *P. acnes* core or ~24,000 SNVs per sample after filtering variants for an allele frequency > 1% and > 4 read depth. To test if *P. acnes* sequencing depth was sufficient to identify the majority of variants, we generated rarefaction curves of SNVs discovered over increasing read depths [Figure 4.3D]. We found that 1 million reads, 40X coverage of the *P. acnes* core, was sufficient for variant discovery.

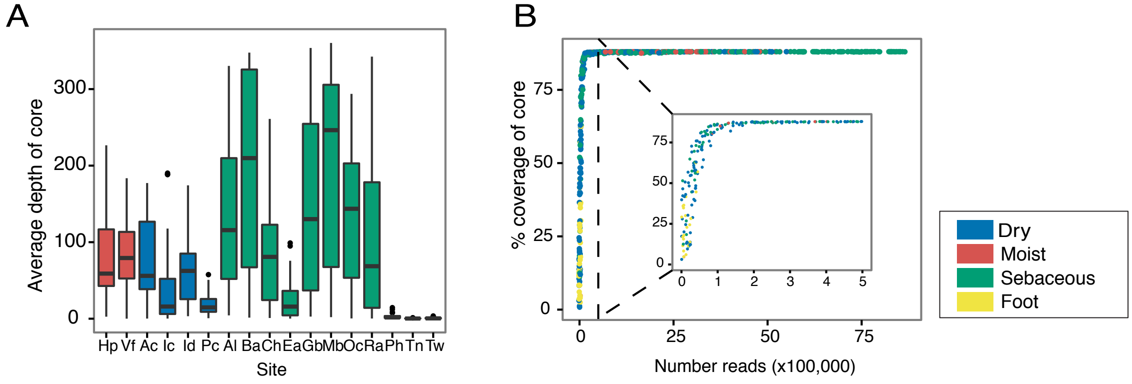


Figure 4.5. *P. acnes* core coverage across body sites.

A) Boxplots show average depth of coverage of the *P. acnes* core across body sites. Black lines indicate median, boxes show first and third quartiles. Colors correspond to the site characteristic. B) Number of *P. acnes* reads versus percent coverage of the *P. acnes* core. With ~100,000 reads, the complete core is covered.

The major advantage of an SNV-based approach is that given sufficient sequencing depth, temporal stability and genetic diversity can be estimated in the absence of a large number of sequenced reference strains. By comparing the sharedness of SNVs between timepoints or individuals using the Jaccard index, which measures similarity based on presence/absence of features, we found the stability of *P. acnes* SNVs mirrors that of *P. acnes* strains over time [Figure 4.3C]. We found that regardless of duration, an individual shares significantly more SNVs with themself over time than with other individuals (P-value<0.001).

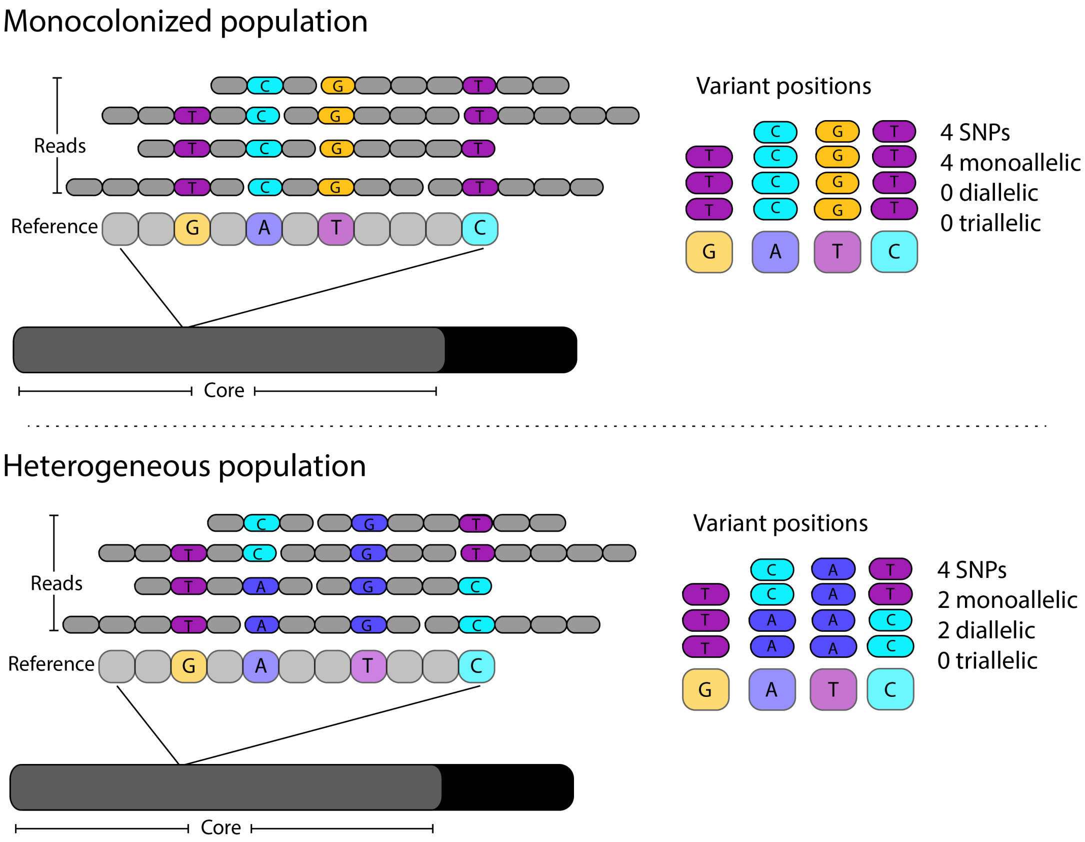


Figure 4.6. SNV distributions indicate heterogeneity of a community.

Top) In a homogenous community, with a single strain per species, when reads from that strain are mapped to a reference, all variant positions will be monoallelic. Bottom) In a heterogeneous community of multiple strains, some variants will be unique to a strain, while other variants are shared between strains. Because of the presence of multiple sources of variation, some variant positions will be monoallelic while others will be di- or in rare cases triallelic.

Finally, polyallelism can reflect genetic heterogeneity, i.e., the presence of multiple strains, in a community [Figure 4.6]. While the number of diallelic sites does not scale linearly with the number of strains in a sample, low levels are indicative of a monocolonized population. We focused on diallelic states, as triallelism was extremely rare (<1.0% in the population). We calculated the cumulative distribution of alternate alleles derived from the shotgun metagenomic reads as a function of distance from a defined reference genome [Figure 4.7]. Through these analyses, we identified a putatively monocolonized individual with strikingly low diallelism (1.0% of sites), in contrast to other individuals, e.g., HV02 and HV09, in which 97.4% and 25.6% of alternate sites have reads reliably mapping to both reference and an alternate allele indicating the presence of multiple strains. For future disease studies, fluctuations in pollyallelism, for example, a dramatic decrease in the number of diallelic positions could indicate emergence of a dominant pathogenic strain.

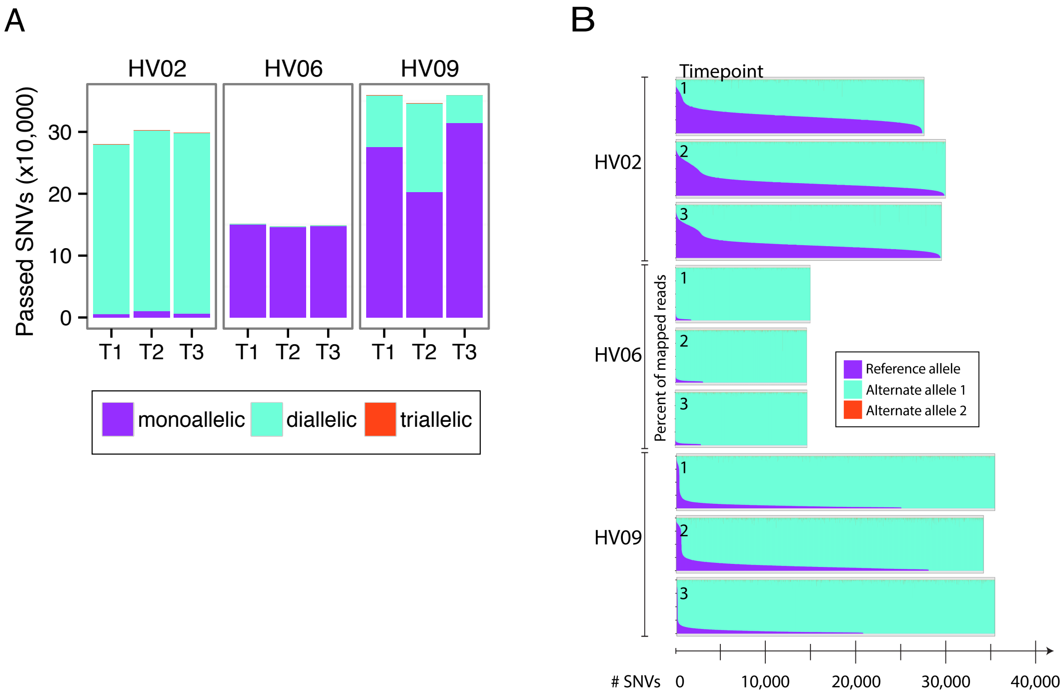


Figure 4.7. *P. acne*s SNV heterogeneity.

A) Number of SNVs that are mono, di, and triallelic. T1, T2, and T3 indicate order in time series. HV is healthy volunteer. B) For samples in [**Figure 4.3**A], the distribution of reads between the reference and alternate allele(s) for all identified SNVs. SNVs ordered by decreasing percentage of reads mapping to the reference allele.

4.3.3 P. acnes pangenome maximized across a multi-phyletic community

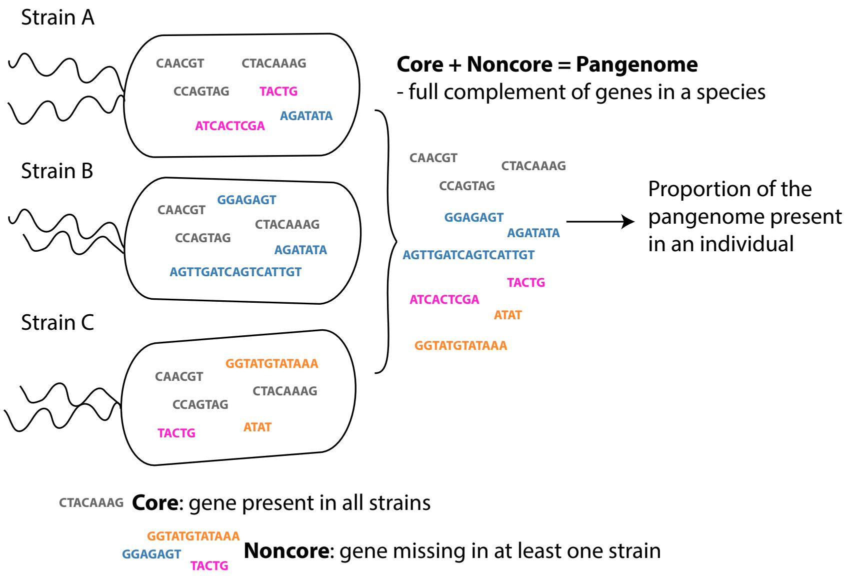


Figure 4.8. Pangenome of a species.

In comparative genomics, the pangenome is the full complement of genes within a species. This example shows three different strains belonging to a single bacterial species. Core genes shown in gray are present in all strains of a species, while noncore genes shown in orange, blue, or pink are present in a subset but not all strains of a species. Together, the core and noncore genes compose the pangenome of a species.

Since skin sites stably maintain the same *P. acnes* strains over time, we wanted to explore the community's full gene content to generate hypotheses about evolutionary forces shaping community drift, resilience, and stability on a functional level. A species' total functional repertoire is the 'pangenome', composed of core (conserved) and non-core (absent in at least one strain) genes [Figure 4.8]. The *P. acnes* pangenome is composed of 3,774 non-redundant gene clusters, of which 1,685 were core [Figure 4.9A,B]. Each additional genome adds 3 novel genes to the total [Figure 4.9C] (Tomida et al., 2013), implying that the majority of *P. acnes* functional capacity is captured within these 78 reference genomes. We mapped our reads to this *P. acnes* pangenome database, requiring 1 million reads for adequate sequencing depth [Figure 4.9D]. Between 82.6 (3,117) and 99.9% (3,771) of the known *P. acnes* pangenome [Figure 4.10A] was represented in healthy individuals. 70% of samples had > 95% (3,585) of the pangenome.

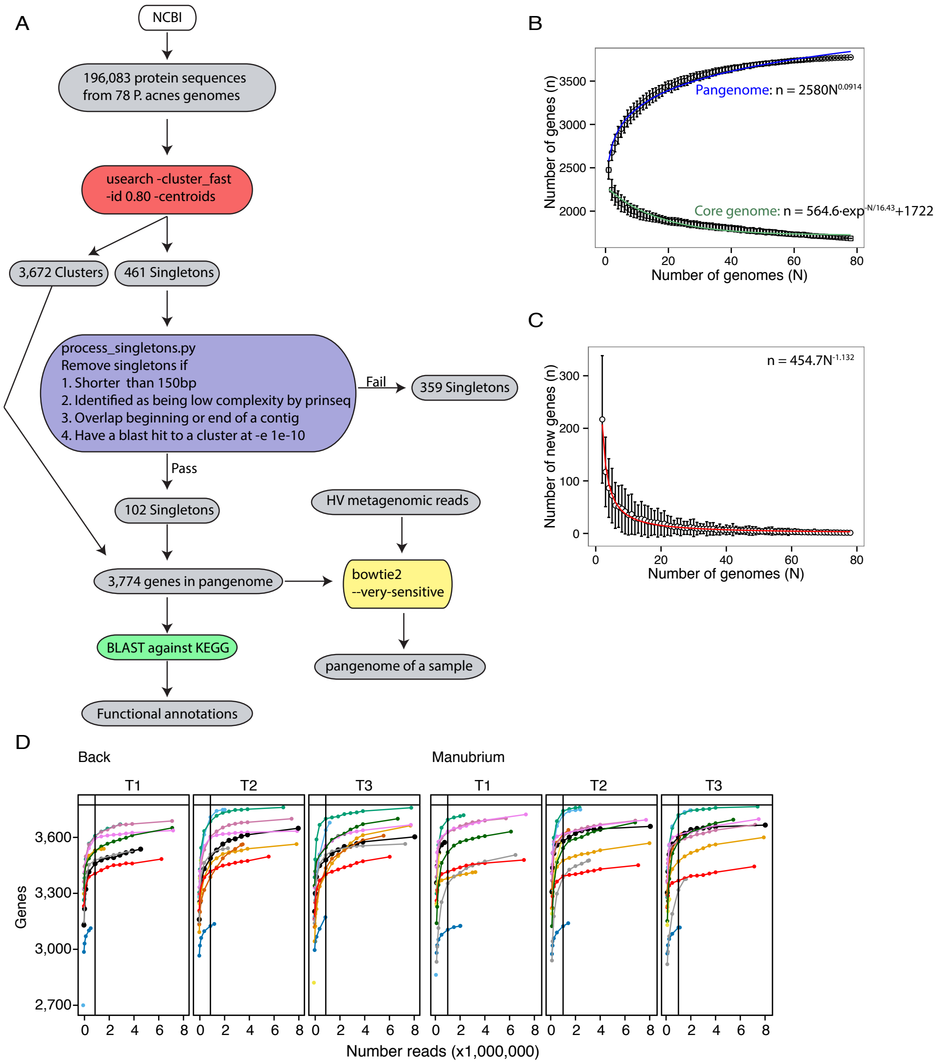


Figure 4.9. Pipeline for *P.* acnes pangenome identification.

A) Detailed pipeline for clustering of the *P. acnes* proteins. For all 78 *P. acnes* genomes, 196,083 protein annotations were downloaded from NCBI. All sequences were clustered with usearch into 3,672 clusters and 461 singletons. Singletons were filtered based on various criteria including length, complexity, and location. After filtering, 102 singletons remained for a total of 3,774 genes in the pangenome. These genes were then BLASTed against a KEGG database to assign functional annotation. Finally, the presence of the gene clusters in the metagenomic samples was determined using bowtie2. B) Gene accumulation curves for pangenome (blue) and core genome (green) as a function of genomes sequences (N). Pangenome data are fit by a power law regression. Core data are fit by an exponential decay curve. Points are means of n for 200 simulations. Error bars indicate the standard deviations for the 200 simulations. C) Accumulation of new genes (n) discovered with the addition of new genome sequences (N) fits a power law regression. D) Rarefaction curves for accumulation of genes for different size subsamples in representative sites. A minimum of 1 million reads was required for further pangenome analyses. Different colors represent trends for different individuals.

Interestingly, combining functional capacity with strain signatures revealed that similar pangenomic capacity can be achieved with distinct strain combinations [Figure 4.10A,B]. This suggests that functional niche saturation can occur through multiple combinations of a limited number of strains, rather than requiring a full phylogenetic complement. Thus, while individuals have distinct *P. acnes* strain signatures, their functional capacities are remarkably similar, only differing 5% between individuals, and an individual is more likely to retain those unique genes over time [Figure 4.10C]. This percent difference between individuals increases to ~40 when relative copy numbers of genes are compared [Figure 4.10D], further illustrating that while individuals' microbial communities have the same set of genes their relative abundances vary.

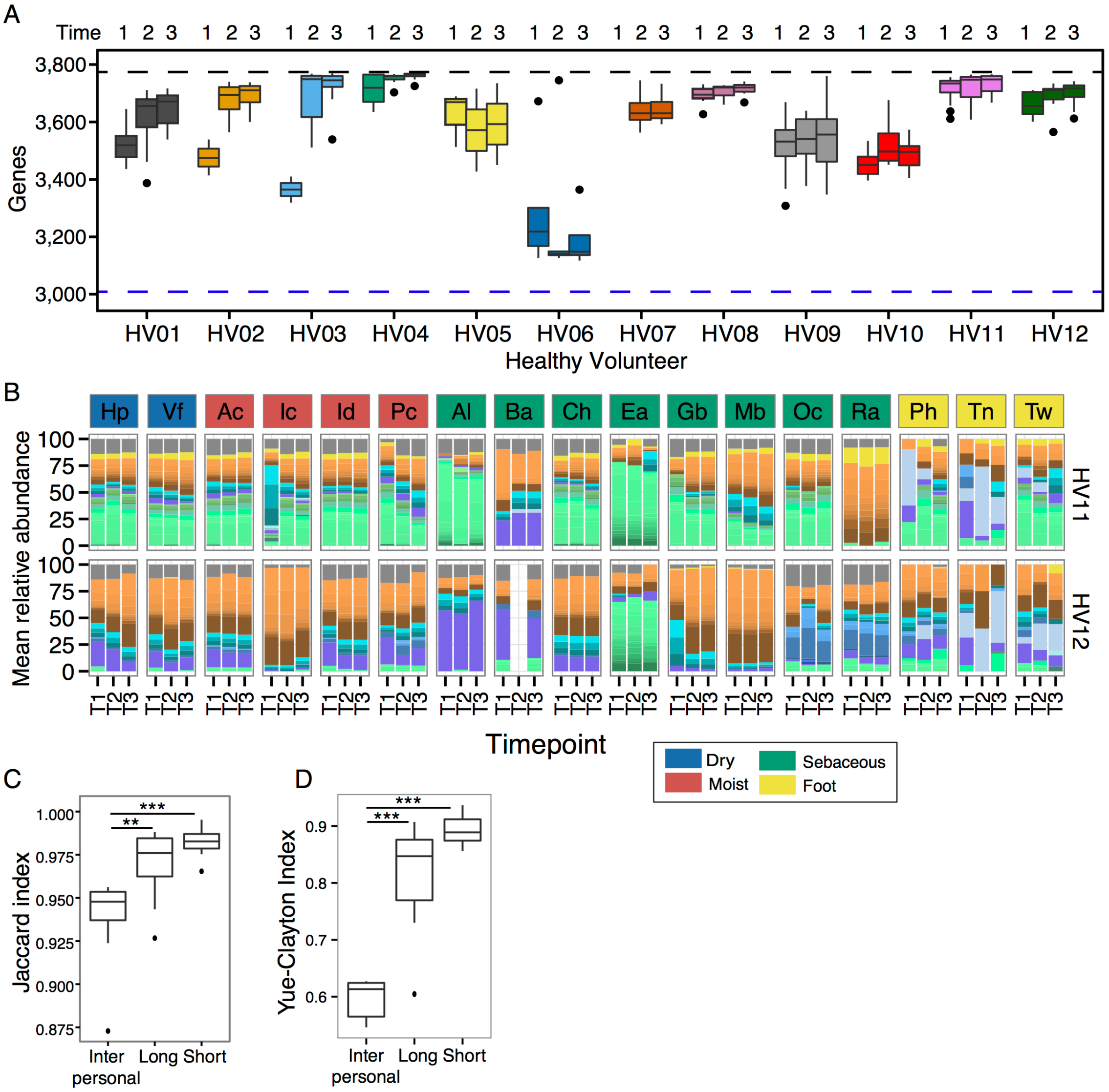


Figure 4.10. *P. acnes* pangenome reaches functional saturation with distinct strain combinations.

A) Boxplots show number of genes present at > 40% coverage across sites of individuals at 3 time points. Black dashed line indicates the 3,774 genes in *P. acnes* pangenome. Blue dashed line indicates the 3,005 genes present in every individual at every time point in at least 50% of body sites. Sites with < 1 million *P. acnes* reads were excluded. B) Relative abundance plots of *P. acnes* strains across all body sites, color-coded by site characteristic. Strain colors defined in **Figure 4.4**B. T1, T2, and T3 indicate order in time series. C) Boxplots of Jaccard theta indices indicate stability of gene presence over time (θ = 1 is identical). ***P* <0 .01, ****P* < 0.001, Wilcoxon rank-sum test. D) Boxplots of Yue-Clayton theta indices indicate the stability of gene copy numbers over time (θ = 1 means identical).

In addition to gene retention, we also observed an increase in pangenome size in three of the individuals (HV01, HV02, and HV03) over time [Figure 4.10A]. Thus, although strain stability is more typical, individuals can acquire new gene content over time. Because of convergent functionality between individuals, we redefined the core genome to represent a ‘functional’ core that is characteristic of healthy communities, towards which a probiotic approach might strive. We defined that 2,982 genes were present in at least 50% of sites in all individuals at all times, exceeding the 1,860 genes that are derived using genomes alone. This functional core genome increases to 3,186 genes when we exclude HV06 whose strain community is predominated by a single *P. acnes* clade [Figure 4.4].

To evaluate functional enrichment within the 769 genes not identified as core, we assigned KEGG pathway annotations to clusters with BLAST. Unsurprisingly, these noncore genes were statistically enriched for pathway "None" (702 of 769 genes) [Figure 4.11A], underscoring that more extensive gene annotations are needed to better understand functional variation. After removing unannotated clusters, we found accessory genes to be enriched in functions associated with ABC transporters and cysteine and methionine metabolism [Figure 4.11B]. Ubiquitous across bacteria, ABC transporters facilitate communication between bacteria and the environment through the active transport of substances such as ions, sugars, lipids, proteins, and drugs across membranes, contributing to nutrient sensing and other processes.

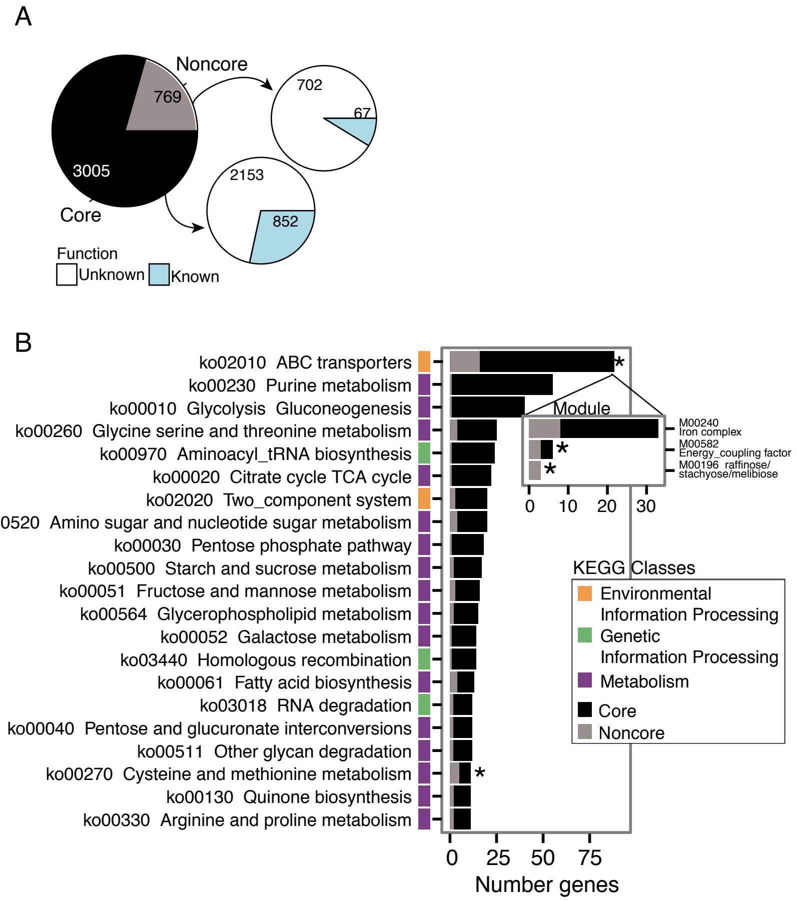


Figure 4.11. ***P. acnes*** **pangenome KEGG functions.**

A) Pie chart indicates distribution of *P. acnes* genes between those present in all individuals “core” and those absent from some individuals “noncore”. Majority of core and noncore genes are pathway “unknown” when compared to a KEGG database. B) Distribution of core and noncore genes for prevalent KEGG pathways. Colors indicate broader KEGG class of each pathway. Pathways with * are functionally enriched in noncore based on Fisher exact test with FDR <0.05.

4.3.4 Multi-phyletic S. epidermidis communities have more variable gene content

To determine if our observations in *P. acnes* extended to other members of the skin community, we applied our analyses to evaluate the stability of S. *epidermidis* strains over time. Like *P. acnes*, *S. epidermidis* is a common skin commensal with well-documented sequence and gene content variation. Multi-phyletic communities of *S.*

epidermidis strains are stably maintained over time irrespective of body site Figure 4.12A, Figure 4.13] strain similarity over the long- and short-term exceeded similarity between individuals (*P*-value< 0.001) [Figure 4.12B], suggesting that new strains are rarely acquired from outside sources. Because *S. epidermidis* is maintained at an overall lower abundance on the skin than *P. acnes* (<10% versus 40%), we lacked sufficient depth for SNV analyses.

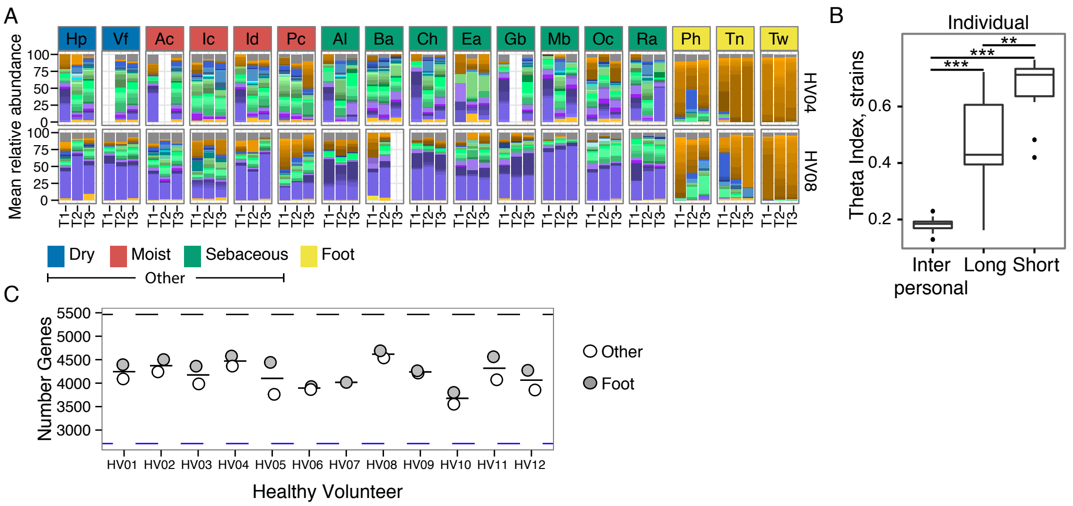


Figure 4.12. *S. epidermidis* strains remain stable over time, and communities do not reach gene saturation.

A) Relative abundance plots of *S. epidermidis* strains across body sites, color-coded by site characteristic. Full set of taxonomic classifications is shown in Figure S6B. Strain colors defined in Figure **Figure 4.13**A. Due to coverage *S. epidermidis*, foot sites were combined as “foot” and all moist, dry, and sebaceous sites were combined as “other”. Combined samples with <1 million *S. epidermidis* reads were excluded. B) Boxplots of theta indices indicate the stability of *S. epidermidis* strains within an individual over long or short time compared to between individuals (Inter) (θ = 1 means identical). ***P* < 0.01; ****P* < 0.001, Wilcoxon rank-sum test. C) Number of genes present on foot and other body sites, with mean number shown as thin bar. Black dashed line indicates the 5,465 genes present in *S. epidermidis* pangenome. Blue dashed lines indicate the 2,712 genes present in all foot and other combined samples.

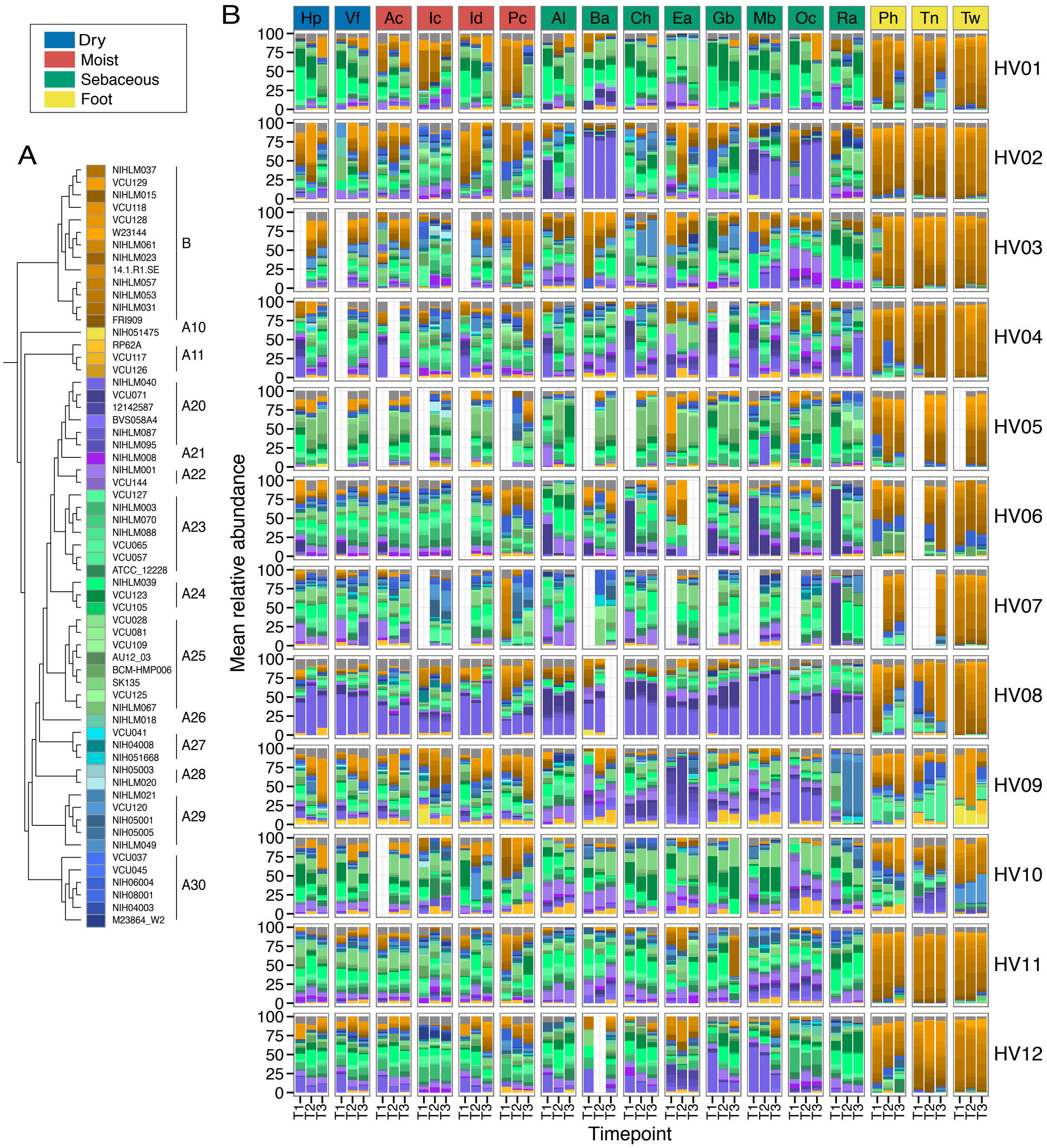


Figure 4.13. ***S. epidermidis*** **strains are stable over time across body sites.**

A) Dendrogram of *S. epidermidis* strain similarity based on core SNVs. Similar strains are grouped into clades. B) Full *S. epidermidis* strain tracking for all individuals, all sites. Sample headings are colored by site characteristic. T1, T2, and T3 indicate timepoints, with T1 and T2 being long-duration timepoints (>1 year) and T2 and T3 short-duration timepoints (~1 month). Colors correspond to those in (A).

The *S. epidermidis* pangenome, assembled from 61 reference genomes, is larger than *P.acnes*, containing 5,465 unique clusters and a noncore of 3,583 genes [Figure 4.14A] with 23 genes added from each additional genome [Figure 4.14B]. To achieve sufficient coverage for pangenome analyses, S. *epidermidis* reads from three foot sites were pooled by individual to create 'foot', while non-foot body sites were pooled to yield a composite "other" sample [Figure 4.14C]. This grouping was based on B clade S. *epidermidis* dominance on all foot sites [Figure 4.13]. Using samples with greater than 1 million S. *epidermidis* reads [Figure 4.14D], we found that individuals did not appear to reach gene saturation, and generally possessed from 65 to 85% (3,552 to 4,693 genes) of the available pangenome [Figure 4.12C]. In addition to lower saturation of the pan-genome, *S.epidermidis* gene content varied 20% between individuals, in contrast to 5% for *P. acnes*. Accordingly, our newly defined *S. epidermidis* healthy 'core' is 2,712 genes. Unique genes could encode similar functions, but be unrecognized as homologs, which could increase the number of genes in the *S.epidermidis* pan-genome, leading to lower full representation. However, the differences in the functional/gene saturation may also be explained by the relatively narrow niche of the sebaceous gland where *P. acnes* primarily resides. *S. epidermidis* has a broader range, which could be reflected in the larger complement of genes needed for strains to persist in a niche.

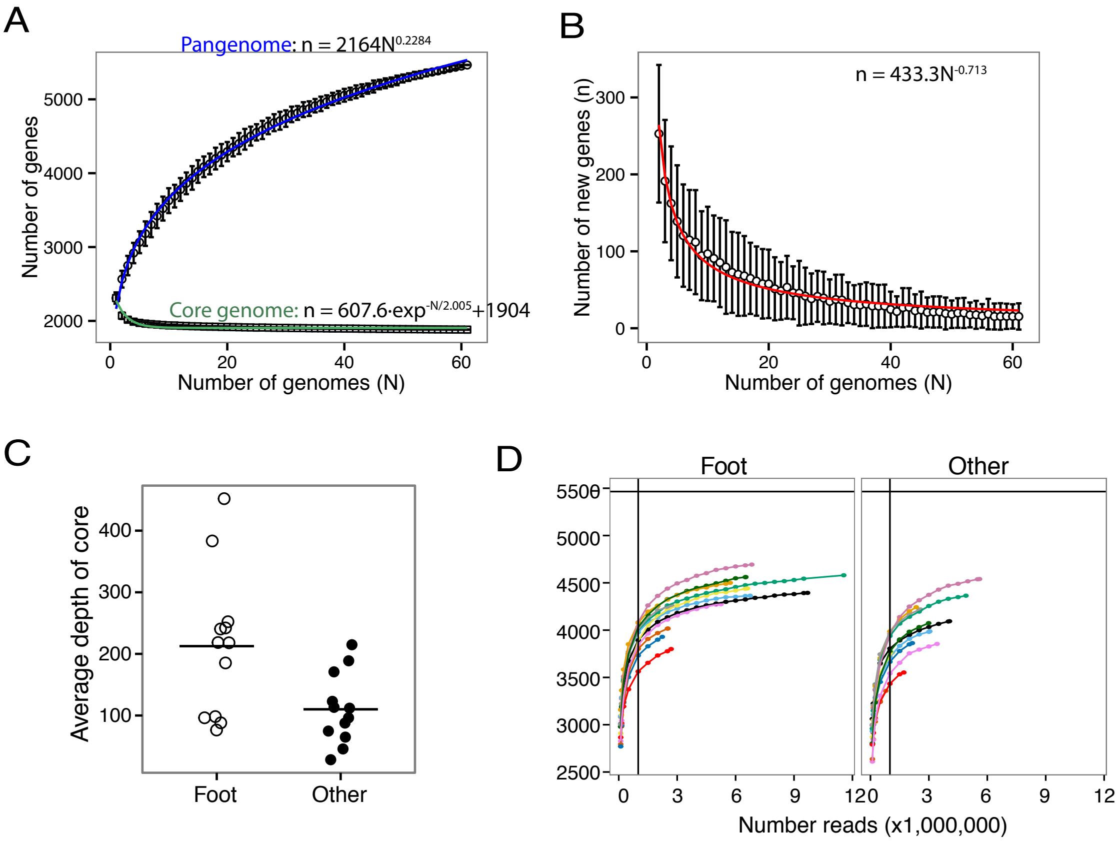


Figure 4.14. ***S. epidermidis*** **pangenome assembly.**

A) Gene accumulation curves for pangenome (blue) and core genome (green) as a function of genomes sequences (N). Pangenome data are fit by a power law regression. Core data are fit by an exponential decay curve. Points are means of n for 200 simulations. Error bars indicate the standard deviations for the 200 simulations. B) Accumulation of new genes (n) discovered with the addition of new genome sequences (N) fits a power law regression. C) Due to low *S. epidermidis* coverage across samples, all foot sites of an individual were combined as "foot" and all moist, dry, and sebaceous sites were combined as "other". Average depth of the *S. epidermidis* core for each sample is shown. D) Rarefaction curves for accumulation of genes for different size subsamples in the combined foot and other samples. A minimum of 1 million reads was required for further pangenome analyses. Different colors represent trends for different individuals.

To examine the functional diversity of the healthy noncore, we examined KEGG annotations for enrichment. Similar to *P. acnes*, the noncore was statistically enriched for pathway "None", with only 139 annotations for 2,753 genes [Figure 4.15A]. Of annotated clusters, we identified DNA replication and carotenoid biosynthesis as enriched within the noncore [Figure 4.15B]. Other enriched pathways demonstrated significant trends, including beta-lactam resistance, cysteine and methionine metabolism, bacterial secretion system, vancomycin resistance, and steroid hormone biosynthesis. Discovering multiple mechanisms of drug resistance in the noncore was unsurprising, given that individuals have varied histories of antibiotic usage, and intraspecies transfer of drug resistance is common. Finally, we looked for functional differences in the foot compared to non-foot sites and found overall gene content was not statistically significantly different, despite the presence of 212 genes specific to foot and 58 genes to non-foot sites.

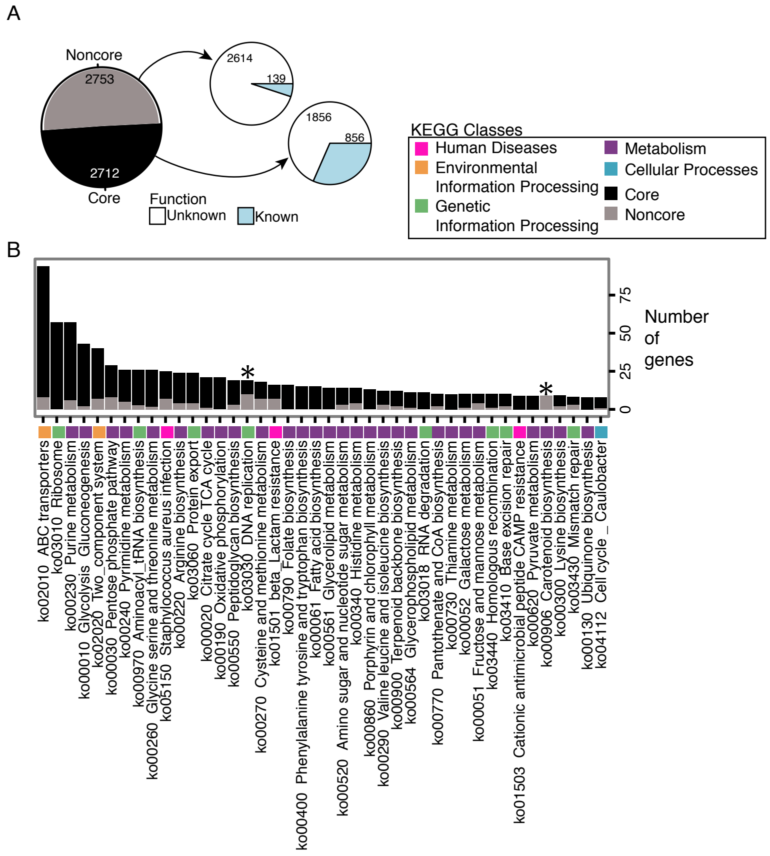


Figure 4.15. *S. epidermidis* KEGG pangenome function.

A) Pie chart indicates distribution of *S. epidermidis* genes between those present in all individuals "core" and those missing from some individuals "noncore". Majority of core and noncore genes are pathway "unknown" when compared to a KEGG database. B) Distribution of genes between core and noncore for the most prevalent KEGG pathways. Colors indicate the broader KEGG class of each pathway. Pathways indicated with * are functionally enriched in noncore based on Fisher exact test with FDR < 0.05.

4.3.5 Functional differences between S. epidermidis strains

To test the functional consequences of genomic variability between *S. epidermidis* strains, 4 isolates from distinct clades were tested in a murine model of microbial association [Figure 4.16B](Naik et al., 2015; Naik et al., 2012). In this model, overnight cultures of bacteria were topically applied to the skin of wild type mice with intact epidermal barriers and immune systems. This was done every other day four times [Figure 4.16A]. On day 14, the mice were sacrificed for analysis and their cutaneous immune cells were analyzed with flow cytometry, a technique that allows proteins on the surface and inside of cells to be tagged with fluorescently labeled antibodies. With this approach, individual cells can be classified in a mixed population. On day 14 post topical association, analysis of immune cells in the mouse ears revealed that different *S. epidermidis* isolates could induce variable adaptive immune responses. In mice associated with 3 of the 4 isolates, there was an increased accumulation of T cell receptor (TCR) $\alpha\beta^+$ cells that were $CD4^+$, while only mice associated with the A20 isolate showed a large increase in the number of $CD8^+$ T cells [Figure 4.16C,D,E]. In contrast, the A25 isolate induced minimal immune responses compared to controls in this model. Differences were also observed in the effector potential of these cells. $TCR\beta^+$ cells in the A20 associated animals had greater potential to produce the cytokine interferon gamma (IFNg) [Figure 4.16F].

Further experiments are necessary to determine whether genetic elements or expression differences are conferring these functional variations. Additionally, because individuals harbor heterogeneous communities of *S. epidermidis* strains, other experiments should be done to monitor what happens when multiple strains are applied simultaneously. When applied as communities, the strains could actively inhibit one another or act synergistically to amplify an effect. Such experiments will be necessary to begin to explain how heterogeneous communities are assembled and why they are stably maintained over time.

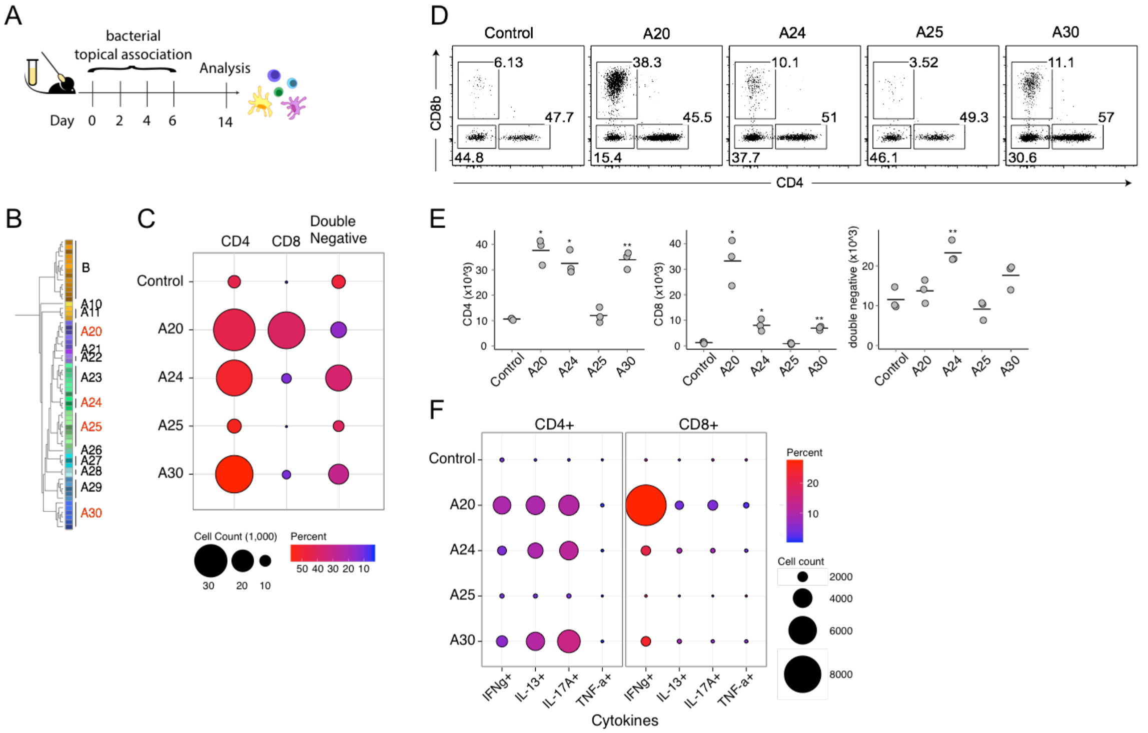


Figure 4.16. *S. epidermidis* strains induce unique immunological signatures.

A) Experimental study design B) *S. epidermidis* strain cladogram with clades applied to mice highlighted in red C. TCRβ$^+$ subpopulation summary plot from topical application of strains in A. Circle size corresponds to absolute numbers of cells while color represents percentage of the TCRβ$^+$population. D) Representative flow plots for C. E) Absolute numbers for C. F) Summary plot of cytokine production from CD8$^+$ and CD4$^+$ T cell populations. Circle size corresponds to absolute numbers of cells while color represents percentage of the CD4$^+$ or CD8+ population.

4.4 Discussion

Despite the continuous perturbation that human skin undergoes in daily life, healthy adults stably maintain their skin communities for up to two years, similar to the stability observed in the gut (Faith et al., 2013; Schloissnig et al., 2013). Homeostasis of skin microbial communities is largely maintained by fixation of abundant species, although a smaller number of low abundance species are also stably maintained and contribute to an individual's unique microbial signature. We suspect that larger, longer-term studies will show a larger reservoir of transients entering and exiting the community, consistent with previous observations of individuals sharing and receiving microbes from the home and other individuals (Lax et al., 2014). Such stochastic drift likely increases over time, unless other constraints, like geographic restriction, lifestyle, or host immune surveillance narrow the transient pool.

We surmise that in the absence of major perturbations, dominant characteristics of skin microbial communities would remain stable indefinitely, a conclusion previously extrapolated for gut communities (Faith et al., 2013; Schloissnig et al., 2013). This stability extends beyond the species level into SNVs and strains, which can impart unique functional contributions to a niche or individual. Total functional content variation, however, differed depending on skin species—*P. acnes,* a dominant skin commensal, showed low content variation in comparison to *S. epidermidis*. However, integration of metagenomic, or 'coding' potential with transcriptional or metabolomics profiles may better delineate community function, as SNVs and small variants can impact actual functional levels.

Future studies will define what and how extrinsic perturbations can alter the skin microbiota; these include antimicrobial treatment (Naik et al., 2012), probiotics, prebiotics, long-term environmental relocations, or diet (Kang et al., 2015). Intrinsic conditions, like immunosuppression, illness, or the occurrence of disease, have also been shown to cause major shifts in skin communities (Kong et al., 2012; Oh et al., 2013). In future disease studies, sequence data can generate hypotheses about which strains contribute to the disease and which are bystanders in the greater microbial consortia. Subsequently, valuable functional information can be gained from culturing and sequencing of primary isolates associated with metagenomic datasets. Functional assaying of individual and mixed strain groups *in vitro* and in animal models will be particularly relevant for determining the causality of diseases. Such studies are the prelude to prebiotic, probiotic and transplantation approaches of skin microbes in the context of disease amelioration and prevention.

4.5 Materials and Methods

4.5.1 Subject recruitment and sampling

To expand our previous metagenomic survey, we re-sampled 12 healthy volunteers from our original study (Oh et al., 2014). Recruitment criteria, sampling procedure, and sample processing were as described previously. Briefly, 7 males and 5 females adults <45 years without chronic skin diseases were sampled three times between June 2011 and May 2014. Sample collection was approved by the Institutional Review

Board of the National Human Genome Research Institute (http://www.clinicaltrials.gov/ct2/show/NCT00605878) and all subjects provided informed consent. Longitudinal samples were collected such that the span between time 1 and time 2 was 10-30 months, while 5-10 weeks separated time 2 and time 3 [Figure 4.2A]. This study design allowed the comparison of stability over a long and short time span. Individuals with a history of chronic medical conditions, including chronic dermatologic diseases, were excluded. 3 patients did report use of oral antibiotics between timepoint 1 and timepoint 2. However, in this study, antibiotic usage did not appear to induce discernible shifts in the overall diversity or structures of skin communities. Separate studies are necessary to fully understand the effects of oral antibiotics on the skin.

17 sites were sampled to represent the diverse physiological characteristics of skin and the sites of predilection for certain dermatologic disease [Figure 3.5]: dry (hypothenar palm, volar forearm), moist (antecubital crease, inguinal crease, interdigital web space, popliteal crease), sebaceous (alar crease, back, cheek, external auditory canal, glabella, manubrium, occiput, retroauricular crease), and foot (plantar heel, toenail, toe web space). To obtain sufficient DNA for metagenomic sequencing, most sites were sampled using a swab-scrape-swab procedure, exceptions include the external auditory canal where only a swab was used and the toenail where a clipping was taken. All samples were stored in lysis buffer at -80C until DNA extraction.

4.5.2 Sample sequencing

Procedures for library generation, sequencing, and processing of longitudinal samples were as previously described (Oh et al., 2014). Briefly, Nextera library kits were used to generate Illumina libraries per manufacturer's instructions with the exception of increasing from 6 to 10 PCR cycles. Libraries were sequenced on an Illumina HiSeq at the NIH Intramural Sequencing Center to a target of 15 to 50 million clusters of 2 x 100bp reads. In total, for 12 individuals, 3 timepoints, we obtained 594 samples or 8.4 trillon reads (722 Gbp) of non-human, quality-filtered paired-end and singleton reads (median 17.9 million reads (1.4 Gbp) per sample). After human removal based on mapping to the hg19 human reference genome, all samples were processed to trim bases with quality score below 20 and remove reads less than 50 bp. To reduce computational burden, post quality control, samples with >20 million reads were subsampled to 10 million paired end reads, and singletons were discarded.

4.5.3 *Taxonomic classifications of skin species and diversity estimates*

Taxonomic classifications were performed as previously described (Oh et al., 2014), except we updated the viral database, incorporating all Refseq viral genomes as of 06.2015. The microbial reference genome database in total included 2342 bacterial, 389 fungal, 6009 viral, and 67 archaeal. Reads not matching hg19 + hg19 rRNA were mapped to this genome collection using bowtie2’s —very-sensitive parameter retrieving the top 10 hits (Langmead and Salzberg, 2012). Reads mapping to multiple genomes were then reassigned using Pathoscope v1.0 (Francis et al., 2013), which uses a Bayesian framework to examine each read’s sequence and mapping quality within the context of a global reassignment. Read hit counts were then normalized by genome length and scaled to sum to one. Coverages were calculated using the genomeCoverageBed tool in the Bedtools suite (Quinlan and Hall, 2010). Because very low abundance organisms are represented by few reads, they are more susceptible to misclassification than more abundant genomes. To reduce the effects of low abundance misclassifications, we used genome coverage cutoffs for relative abundance and diversity calculations; genomes were binned with coverage cutoffs of ≥ 1, 0.1, 0.01 or 0.001. A coverage cutoff of ≥ 1 was used for major analyses, a conservative number that produced classifications that most closely corresponded with the results from other common metagenomic classifiers (e.g., Metaphlan (Truong et al., 2015) or analysis using other methodologies like 16S rRNA and ITS gene sequencing (Oh et al., 2014). This number typically accounts for >99.9% of the community abundance. All taxonomies were reconstructed to the species level, combining hits to multiple strain subtypes to reduce the potential for erroneous strain-calling.

4.5.4 Strain tracking of dominant species

Strain tracking of the dominant skin commensals *Propionibacterium acnes* and *Staphylococcus epidermidis* was accomplished as described previously (Oh et al., 2014). Briefly, reference databases for *P. acnes* and *S. epidermidis* were compiled from all complete and draft genomes available on NCBI, 78 and 61, respectively. Whole genome alignment, with nucmer, was then used to identity the "core" region shared between all sequenced strains for a species. SNVs identified in these core regions were subsequently used to generate dendograms with PhyML 3.0. We then grouped strains into subtypes based on phylogenic distance, 12 for *P. acnes* and 14 for *S. epidermidis* [Figure 4.4B, Figure 4.13A]. Metagenomic reads were mapped to each species database with bowtie2 (-score-min L,-0.6,0.006, -k number of genomes) (Langmead and Salzberg, 2012) with zero tolerance for mismatches. The resulting alignment file was then processed with Pathoscope (-theta_prior 10 x 10^88) (Francis et al., 2013) to deconvolute multiple mapping reads. Accuracy of this strain-tracking approach was previously validated with extensive simulations (Oh et al., 2014).

4.5.5 Identification of SNVs in the P. acnes core

For each sample, coverage of the *P. acnes* core was calculated with samtools (Li, 2011) and genomecoveragebed (Quinlan and Hall, 2010). High average coverage nicely related to percent coverage of the *P. acnes* core [Figure 4.5]. Back and manubrium samples had the highest *P. acnes* sequencing depth, so were selected for more extensive SNV analysis [Figure 4.5]. Because *P. acnes* strains are shared across sites of an individual, these results can be extrapolated to the rest of the body. For SNV analysis [Figure 4.17], metagenomic reads were first mapped against the *P. acnes* core genome using bowtie2 (--very-sensitive). The resulting alignment file was sorted by samtools and then processed with GATK's IndelRealigner (McKenna et al., 2010) to minimize mismatches resulting from insertions or deletions in the reads with respect to the reference genome. The corrected alignment file was then analyzed with samtools and bcftools to identify possible variants (samtools mpileup -uD -q30 -Q30, bcftools view -Abvcg, vcfutils.pl varFilter -D99percentileofcoverage -d4 -1 .00001 -4 .00001). Parameters were selected to filter false positive polymorphisms that were a result of sequencing error, recent sequence duplications not found in the draft genome, strand bias, or end distance bias. Possible variants were then filtered with custom scripts to meet criteria previously described (Lieberman et al., 2014). Briefly, an alternate allele was only considered if it was supported by >2 reads with a minimum mapping quality of 30, had an allele frequency >3%, and fewer than 20% of reads supporting the SNV also mapped to an indel. With rarefaction curves of SNVs discovered over increasing read depths [Figure 4.3D], we found that 1 million reads, 40X coverage of the *P. acnes* core, was sufficient for variant discovery. Thus, to reduce computational burden only subsamples of 1 million reads were used for further analysis.

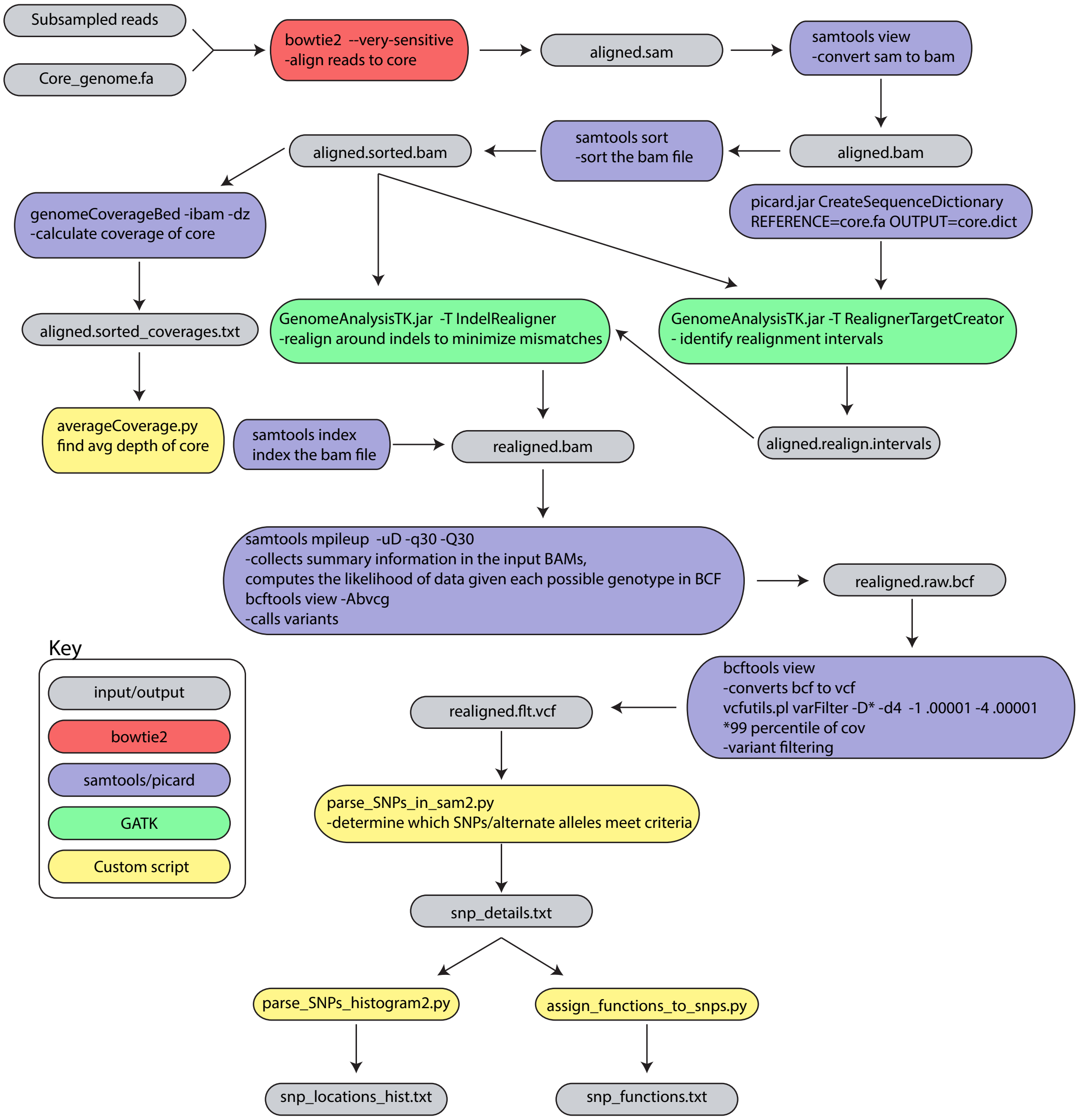

Subsampled reads
Core_genome.fa
bowtie2 --very-sensitive
-align reads to core
aligned.sam
samtools view
-convert sam to bam
aligned.bam
samtools sort
-sort the bam file
aligned.sorted.bam
genomeCoverageBed -ibam -dz
-calculate coverage of core
aligned.sorted_coverages.txt
averageCoverage.py
find avg depth of core
picard.jar CreateSequenceDictionary
REFERENCE=core.fa OUTPUT=core.dict
GenomeAnalysisTK.jar -T RealignerTargetCreator
- identify realignment intervals
aligned.realign.intervals
GenomeAnalysisTK.jar -T IndelRealigner
-realign around indels to minimize mismatches
realigned.bam
samtools index
index the bam file
samtools mpileup -uD -q30 -Q30
-collects summary information in the input BAMs,
computes the likelihood of data given each possible genotype in BCF
bcftools view -Abvcg
-calls variants
realigned.raw.bcf
bcftools view
-converts bcf to vcf
vcfutils.pl varFilter -D* -d4 -1 .00001 -4 .00001
*99 percentile of cov
-variant filtering
realigned.flt.vcf
parse_SNPs_in_sam2.py
-determine which SNPs/alternate alleles meet criteria
snp_details.txt
parse_SNPs_histogram2.py
assign_functions_to_snps.py
snp_locations_hist.txt
snp_functions.txt
Key
input/output
bowtie2
samtools/picard
GATK
Custom script

Figure 4.17. Detailed pipeline for identifying a species SNVs in metagenomic sequencing data.

First, metagenomic reads are mapped against the species of interest's core genome. The resulting alignment file is processed with the programs samtools, picard, and GATK to identify the depth of sequencing coverage and correct for misalignments due to indels. The remaining alignment file is then further processed with samtools and custom scripts to identify single nucleotodie variant regions. Box colors correspond to the programs implemented at each step.

4.5.6 Pangenome analyzes of dominant species

To identify the functional capacity of dominant species in our metagenomic samples, we followed the procedure illustrated in Figure 4.9A. First, 196,083 *P. acnes* nucleotide-coding sequences were downloaded from NCBI and 147,257 *S. epidermidis* sequences were extracted from Manatee annotations of the genomes. The IGS Analysis Engine was used for structural and functional annotation of the sequences. (http://ae.igs.umaryland.edu/cgi/index.cgi, Galens et al., 2011). Manatee was used to view annotations (http://manatee.sourceforge.net/). Genes were then clustered into non-redundant orthologs with usearch (-cluster_fast -id 0.80 -centroids) (Edgar, 2010). To validate accuracy of the clustering, we verified the presence of 13 single copy marker genes (Greenblum et al., 2015). Singletons, clusters composed of a single sequence, were then filtered based on previously established criteria (Lefebure and Stanhope, 2007). Briefly, singletons were excluded if they 1) were shorter than 150 nucleotides, 2) were flagged as low complexity by Prinseq (Schmieder and Edwards, 2011), or 3) overlapped the beginning or end of a contig. 4) had a blast hit to a cluster at -e 1e-10. Based on this criteria 359 *P. acnes* and 874 *S. epidermidis* singletons were removed, leaving 3,774 and 5,627 gene clusters respectively. Gene accumulation curves for these clusters mirrored previous pangenome studies for *P. acnes* (Tomida et al., 2013) and *S. epidermidis (*Conlan et al., 2012*).* The curves showed that new genes discovered with additional genomes and the pangenome followed a power law curve, while core genome size fit an exponential decay curve [Figure 4.9B,C, Figure 4.14A,B]. These gene clusters were then annotated by BLASTx against the KEGG database. To identify the functional capacity of a sample, reads were mapped to each of the gene cluster databases using bowtie2 (--very-sensitive). A gene was subsequently considered present only when 40% of its length was covered with reads. This criteria reduces gene calling due to spuriously mapped reads or reads from orthologs of closely related species (Zhu et al., 2015). Average coverage of each gene was calculated with samtools (Li, 2011) and then normalized by the average coverage of 13 single copy marker genes (Greenblum et al., 2015) to yield a copy number estimate.

4.5.7 Statistics

All statistical analyses were performed in the R software. Data are represented as mean ± standard error of the mean unless otherwise indicated. Spearman correlations of non-zero values were used for all correlation coefficients. Site characteristics were treated as separate groups where indicated based on spatial physiological differences between these different body niches (Grice et al., 2009). For all boxplots, black center lines represent the median and box edges the first and third quartiles. The nonparametric Wilcoxon rank-sum test was used to determine statistically significant differences between microbial populations. Unless otherwise indicated, P-values were adjusted for multiple comparisons using the p.adjust function in R using method = “fdr”. Statistical significance was ascribed to an alpha level of the adjusted P-values ≤ 0.05. Similarity between samples was assessed using the Yue–Clayton theta or Jaccard similarity index with relative abundances of species, sub-strains, or shared genomic variants. The theta coefficient assesses the similarity between two samples based on (1) number of features in common between two samples, and (2) their relative abundances with $\theta = 0$ indicating totally dissimilar communities and $\theta = 1$ identical communities (Yue and Clayton 2005). As θ takes into account species abundance, it is less susceptible to low-abundance species whose classifications are less robust. The Jaccard similarity index is a metric defined by the union of the species occurring between two samples. To avoid repeated measures, samples belonging to an individual were averaged before statistical comparisons between site characteristic when using summary metrics such as means, diversity, or theta indices.

CHAPTER 5 Staphylococcal strain diversity underlies the individuality of atopic dermatitis

5.1 Abstract

Staphylococcus aureus has been tightly linked with atopic dermatitis (AD; eczema). We explored microbial temporal dynamics with metagenomic sequencing to investigate the role of staphylococci in AD. Species-level investigation of AD flares demonstrated a microbial dichotomy in which *S. aureus* was predominant on more severely affected patients while *S. epidermidis* was more predominant on less severely affected patients. Metagenomic analyses at the strain-level determined that *S. aureus*-predominant patients were monocolonized with distinct *S. aureus* strains, while all patients had heterogeneous *S. epidermidis* strain communities. To assess the immunologic effects of these species, we topically applied patient-derived strains to mice. AD strains of *S. aureus* were sufficient to elicit skin inflammation associated with infiltration of Th2 and Th17 cells, an immune signature characteristic of AD patients. Integrating sequencing, culturing, and animal models, we explore a model whereby staphylococcal strains contribute to AD progression through activation of the host immune system.

5.2 Introduction

Atopic dermatitis (AD, eczema) is a common inflammatory skin disorder in industrialized countries, affecting 10-30% of children (Eyerich et al., 2015). Pediatric patients suffer from chronic, relapsing, intensely itchy, inflamed skin lesions, and have an increased likelihood of developing asthma and/or hay fever (Bantz et al., 2014). AD is a complex disease in which multiple underlying components, including epidermal barrier impairment, type 2 immunity, and skin microbes, are thought to play a causative role (Eyerich et al., 2015). Over 30 susceptibility loci have been associated with AD, including mutations in the gene encoding the skin barrier protein filaggrin (FLG) (Genetics et al., 2015) and genes linked to the immune system (Palmer et al., 2006).

In addition to host genetics, the relationship between AD and skin bacteria is clinically well recognized. AD skin disease is clinically managed with combinations of antimicrobial approaches (e.g. antibiotics and dilute bleach baths) and anti-inflammatory or immunosuppressive medications (Huang et al., 2009). The efficacy of these antimicrobial treatments is associated with drops in staphylococcal relative abundances (Kong et al., 2012). *Staphylococcus aureus* is commonly cultured from AD skin (Leyden et al., 1974) and murine models have demonstrated exacerbation of eczematous dermatitis with topical application of *S. aureus* (Kobayashi et al., 2015).

With an increasing appreciation of functional differences between strains within a single species, we performed shotgun metagenomic sequencing of AD patient skin to capture the full genetic potential and strain-level differences of the skin microbiome throughout the course of disease. We confirmed an increase of staphylococcal species during disease flares in our cohort and more deeply explored the *S. aureus* and *S. epidermidis* strain diversity of each patient. To test the functional consequence of this strain diversity between patients, we isolated staphylococcal strains from patients and investigated the cutaneous and immunologic effects when applied topically to a mouse model.

5.3 Bacterial communities shift during AD disease progression

To examine the relationship between the skin microbiota and AD, eleven children with moderate-to-severe AD and seven controls (ages 6-12) were recruited to the NIH Clinical Center between June 2012 and March 2015. As AD has a chronic relapsing course, patients were sampled at stable disease state (baseline/B), worsening of disease (flare/F), and 10-14 days after initiation of treatment (post-flare/PF). At each timepoint, disease severity was determined with objective SCORAD (SCORing Atopic Dermatitis), a validated clinical severity assessment tool (Kunz et al., 1997; Oranje et al., 2007; Williams et al., 1994b). Subjects were sampled bilaterally at sites of disease predilection, the inner elbow (antecubital crease/Ac) and behind the knees (popliteal crease/Pc), along with five additional sites to investigate different physiologic skin sites [Figure 5.1]. For baseline samples, subjects refrained from the use of topical medications on sampled sites to reduce potential confounding effects of skin-directed treatments. Five of the eleven AD patients were able to provide a baseline sample, as the others exhibited clinical worsening of skin disease requiring re-initiation of skin treatments. Based on similar bacterial communities observed at baseline and post-flare (Kong et al., 2012), we focused on comparisons between flare and post-flare time points. In total, we sequenced 422 samples, generating 191 Gb of microbial sequence data from 27 AD patient and 7 healthy control visits. During patient flares, AD disease severity was significantly elevated as indicated by higher mean objective SCORAD (38 ± 2.9) as compared to baseline (9.4 ± 1.6, $P < 4.5\text{x}10^{-4}$) and post-flare (11 ± 1.6, $P < 2.8\text{x}10^{-6}$) [Figure 5.2A].

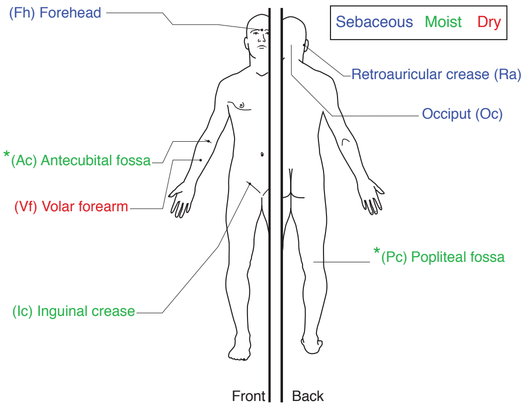


Figure 5.1. Seven sites sampled bilaterally on AD patients and control children.

Sites colored by their microenvironment: sebaceous (blue), moist (green), and dry (red). Sites of AD disease predilection indicated with *.

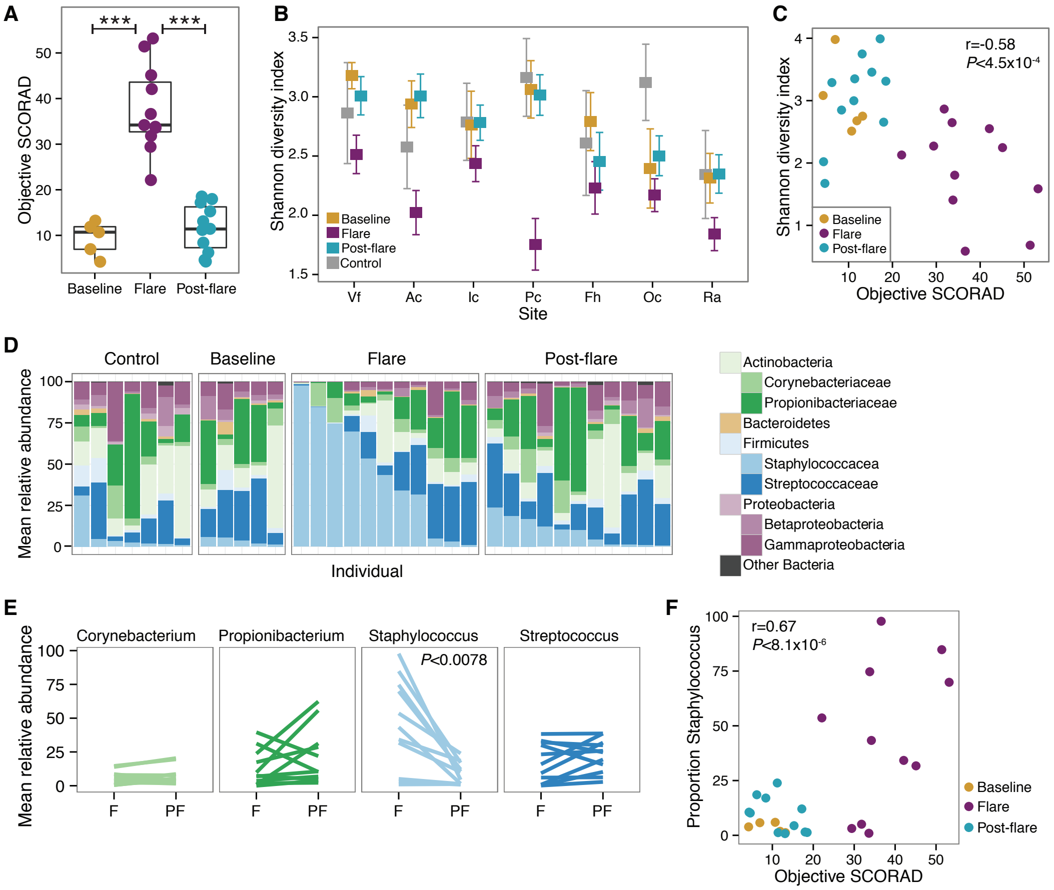


Figure 5.2. Bacterial communities shift during AD disease progression.

A) Objective SCORAD for each patient at baseline, flare, and post-flare. Higher SCORAD corresponds to more severe disease. *** *P*<0.001 B) Mean Shannon diversity +/- SEM in controls and AD disease states. Colors correspond to disease state. Volar forearm (Vf), antecubital crease (Ac), inguinal crease (Ic), popliteal crease (Pc), forehead (Fh), occiput (Oc), and retroauricular crease (Ra). C) Shannon diversity versus objective SCORAD for combined antecubital (Ac) and popliteal creases (Pc) (AcPc) of AD patients. Partial correlation (adjusting for disease state). D) Mean relative abundance of bacterial genera in AcPc for controls and AD disease states. E) Mean relative abundance of predominant genera in AcPc for disease states, Flare (F) and Post-flare (PF). F) Proportion of *Staphylococcus* versus objective SCORAD for AcPc of AD patients, partial correlation (adjusting for disease state).

To compare the microbial community composition across time-points, we mapped microbial reads to a multi-kingdom reference database. In this cohort, we analyzed bacterial communities where we observed the greatest changes [Figure 5.3]; no significant differences in the fungal or viral components over time were identified. We first determined the Shannon diversity index, an ecological measure of richness (total number of bacterial species) and evenness (relative proportion of the bacterial species), to evaluate the overall community structure/composition across body sites and time points. During flares, sites of AD predilection (Ac and Pc) exhibited a marked reduction in Shannon Diversity compared to baseline, post-flare, and healthy controls, a trend observed to a lesser extent across other sites [Figure 5.2B]. Since changes in bacterial diversity were most pronounced at the sites of disease predilection and Ac/Pc have similar microbial communities (Oh et al., 2014), we averaged these sites per subject and used the composite "AcPc" for subsequent analyses. Similar to our previous analysis of microbial diversity in an AD patient cohort (Kong et al., 2012), the partial correlation between objective SCORAD and Shannon diversity, adjusting for disease state, was significantly inversely correlated ($r = -0.58$, $P = 4.5\text{x}10^{-4}$)[Figure 5.2C], indicating that reduced skin bacterial diversity corresponds to worse disease severity, primarily at sites of disease predilection [Figure 5.4A].

To determine which taxa were contributing to the loss of diversity, we compared the relative abundances of the most prominent taxa [Figure 5.2D and Figure 5.4B]. Of the four most prominent genera in the AcPc, only *Staphylococcus* was significantly increased in flares (45 ± 10.2%) as compared to post-flare (9.2 ± 2.4%, $P < 0.0078$) and controls (6.6 ± 4.1%, $P < 0.033$) [Figure 5.2E]. This increase in *Staphylococcus* relative abundance was positively correlated with objective SCORAD (r=0.67,$P < 8.1\text{x}10^{-6}$)[Figure 5.2F], indicating that severe AD was associated with higher staphylococcal relative abundances at sites of disease predilection. In addition, there was a positive correlation for the forehead, retroauricular crease, and volar forearm [Figure 5.4C], sites that can be affected in more severe disease. However, differences in *Corynebacterium*, *Propionibacterium*, and *Streptococcus* relative abundances between flare and post-flare were not significant [Figure 5.2E].

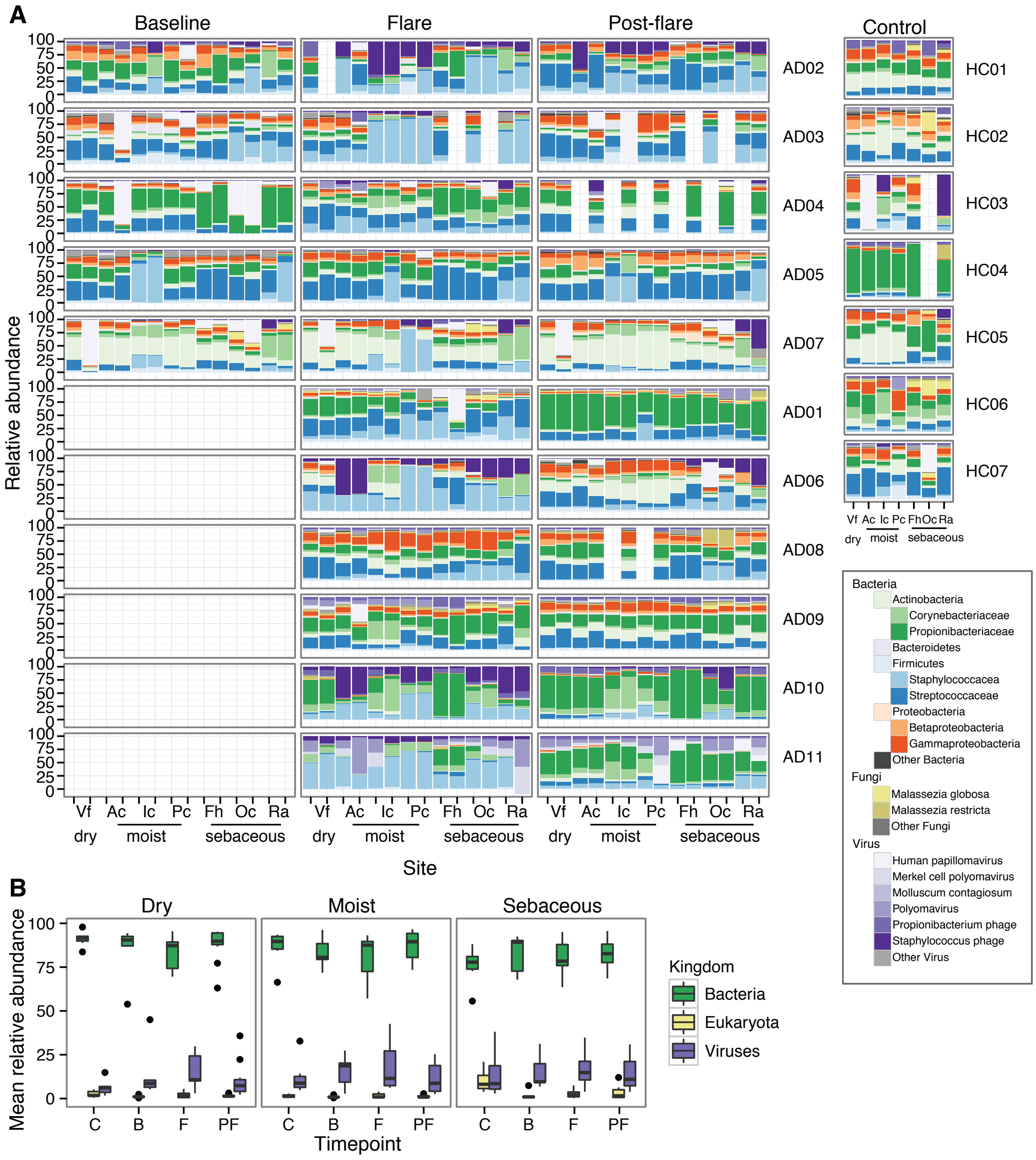


Figure 5.3. Full multi-kingdom taxonomic classifications for AD patients and controls.

A) Relative abundance of most abundant skin taxa for each super-kingdom for all sites in AD patients and controls. B) Boxplots of mean relative abundance of different kingdoms by timepoint for the different site characteristics. Timepoints are Control (C), Baseline (B), Flare (F), and Post-flare (PF).

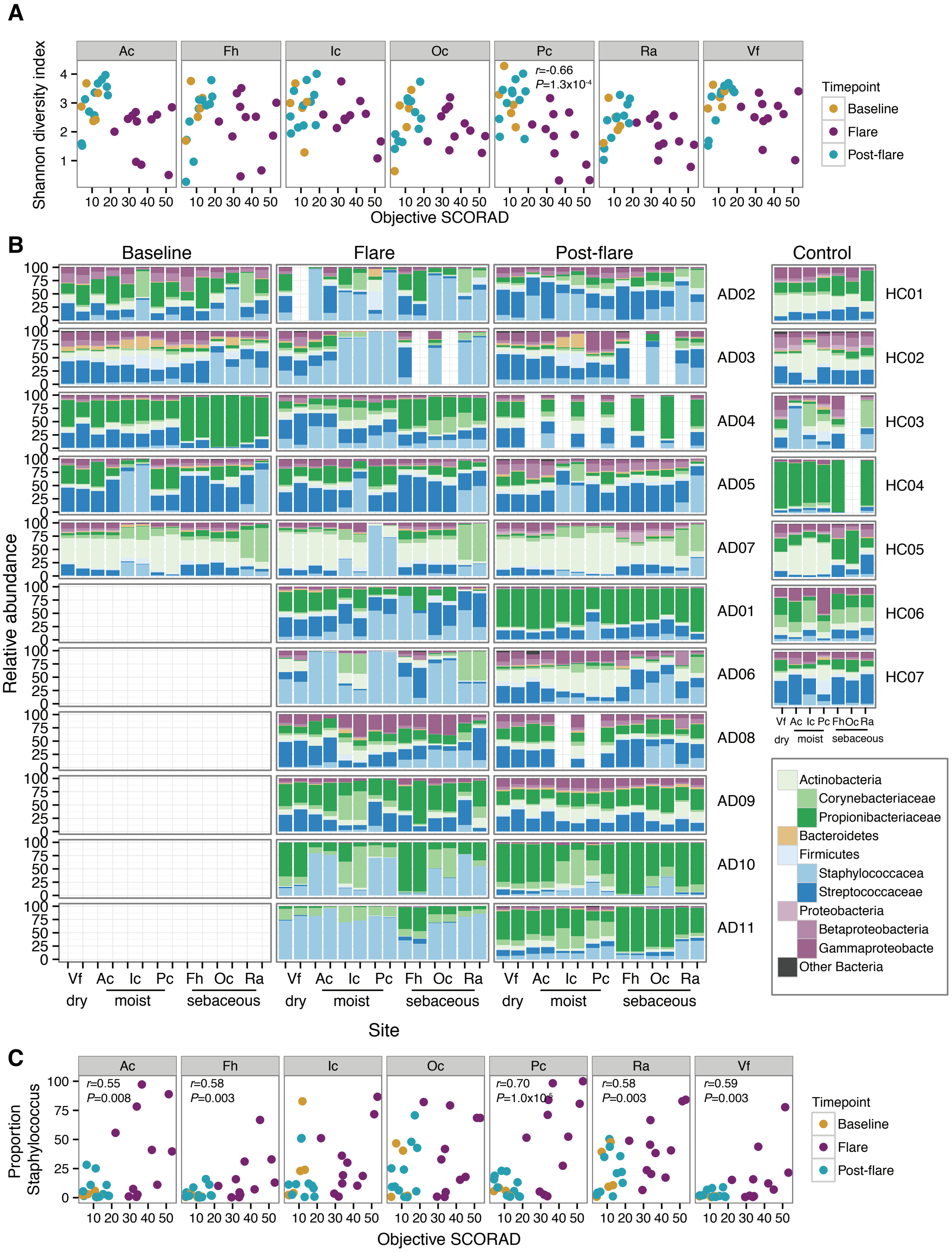


Figure 5.4. Full bacterial taxonomic classifications for AD patients and controls.

A) Shannon diversity versus objective SCORAD for all sites. Partial correlation (adjusting for disease state) with only significant correlations shown. B) Relative abundance of bacterial genera for all sites in AD patients and controls. C). Relative proportion of Staphylococcus versus objective SCORAD for all sites. Partial correlation (adjusting for disease state) with only significant correlations shown.

5.4 More severe AD patient flares associated with specific staphylococcal species

To further examine the positive correlation between *Staphylococcus* and AD disease, (Williams and Gallo, 2015), we identified the relative abundances of staphylococcal species including *S. aureus, S. epidermidis, S. hominis* and *S. capitis* [Figure 5.5A and Figure 5.6]. Only relative abundance of *S. aureus* was significantly decreased from flare (28 ± 8.8%) to post-flare (2.3 ± 0.8%, $P < 0.014$)[Figure 5.5B]. While *S. epidermidis* relative abundances were also higher during flares (13 ± 5.4%) as compared to post-flares (3.7 ± 1.4%), results were not statistically significant. For all patients, relative abundances of *S. aureus* were positively correlated with objective SCORAD ($r = 0.73$, $P < 1.\text{x}10^{-7}$), while *S. epidermidis* was not correlated [Figure 5.5C and Figure 5.7]. This is consistent with the high incidence of *S. aureus*-positive cultures from lesional and nonlesional AD skin (Leyden et al., 1974; Totte et al., 2016). Neither *S. hominis* nor *S. capitis* demonstrated significant shifts in relative abundances between time points [Figure 5.5B] or were correlated with disease severity [Figure 5.7].

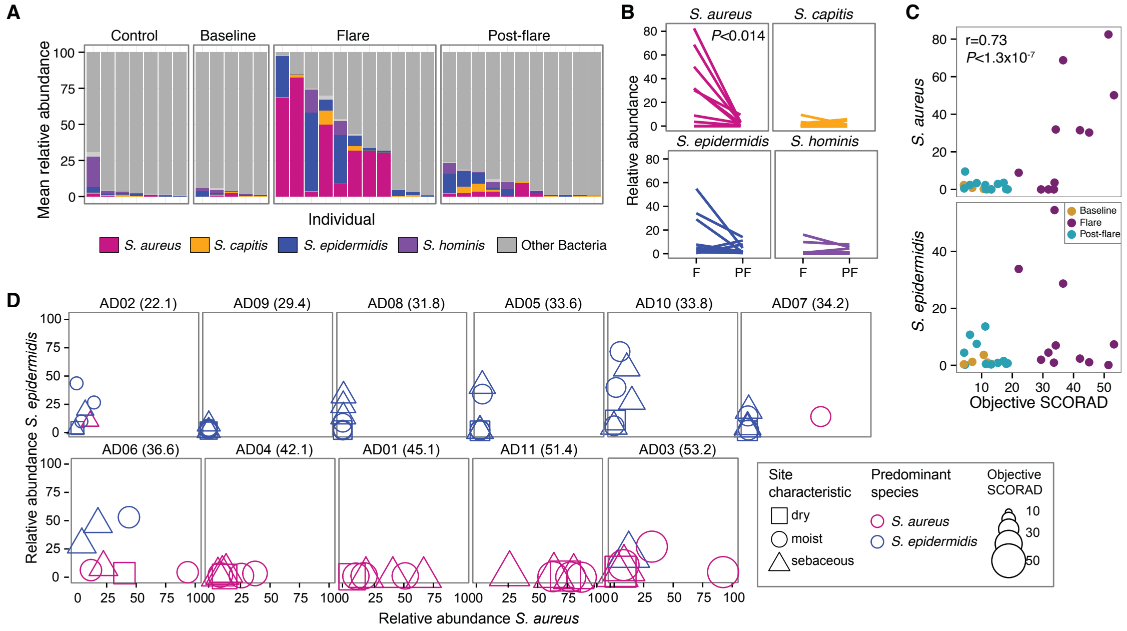


Figure 5.5. Staphylococcal species increase during AD disease flare.

A) Mean relative abundance of staphylococcal species within the total bacterial population in combined antecubital (Ac) and popliteal creases (Pc) (AcPc) of AD patients and controls. B) Mean relative abundance of most abundant Staphylococcus species in AcPc for disease states, Flare (F) and Post-flare (PF). C) Correlation of *S. aureus* (top) and *S. epidermidis* (bottom) mean relative abundance and objective SCORAD for AcPc of patients, partial correlation (adjusting for disease state). D) Comparison of *S. aureus* to *S. epidermidis* relative abundance by patient for all sites. Patient's SCORAD indicated in parenthesis. Shape corresponds to physiological characteristic of the body site, size to the magnitude of disease severity (objective SCORAD), and color to the most predominant species at the site. Patients in the top row have lower SCORADS and a higher predominance of *S. epidermidis* across sites, while bottom row patients are more severe disease and have *S. aureus*-predominantly across sites.

To examine more closely the relationship between staphylococcal species and the degree of disease severity, we plotted the relative abundances of *S. aureus* and S. *epidermidis* based on objective SCORAD [Figure 5.5D]. We observed a trend between patients with more severe AD based on objective SCORAD (45 ± 3.0) and higher relative abundances of *S. aureus* across sampled sites [Figure 5.5D top row]. Patients with less severe disease, or lower objective SCORAD (31 ± 1.9, $P < 0.0043$), had higher relative abundances of *S. epidermidis* across sampled sites [Figure 5.5D bottom row and Figure 5.8]. Based on severity of the objective SCORAD during disease flares, we defined patients as either more severe (34 ± 8.7% *S. aureus*, 7.4 ± 4.2% *S. epidermidis* average across all sites during flare) or less severe (3.8 ± 1.7% *S. aureus*, 13 ± 3.9% *S. epidermidis*). To compare our metagenomics methods with more traditional culture methods, we cultured skin and nares swabs that had been collected concurrently with genomics samples. The sequencing results were more sensitive than the cultivation studies in identifying which patients had *S. aureus*: 100% of patients with more severe disease were culture positive for *S. aureus* at the antecubital crease and nares, while 50% of patients with less severe disease were culture positive for *S. aureus* at those sites. Thus, the close association between AD disease severity and *S. aureus* is even more apparent based on genomics studies. Since sequence-typing methods have been shown to misclassify distinct clones of *S. aureus* as the same clone (Salipante et al., 2015; Ugolotti et al., 2016), a question remained whether a bloom of *S. aureus* observed during disease flares is polyclonal or monoclonal.

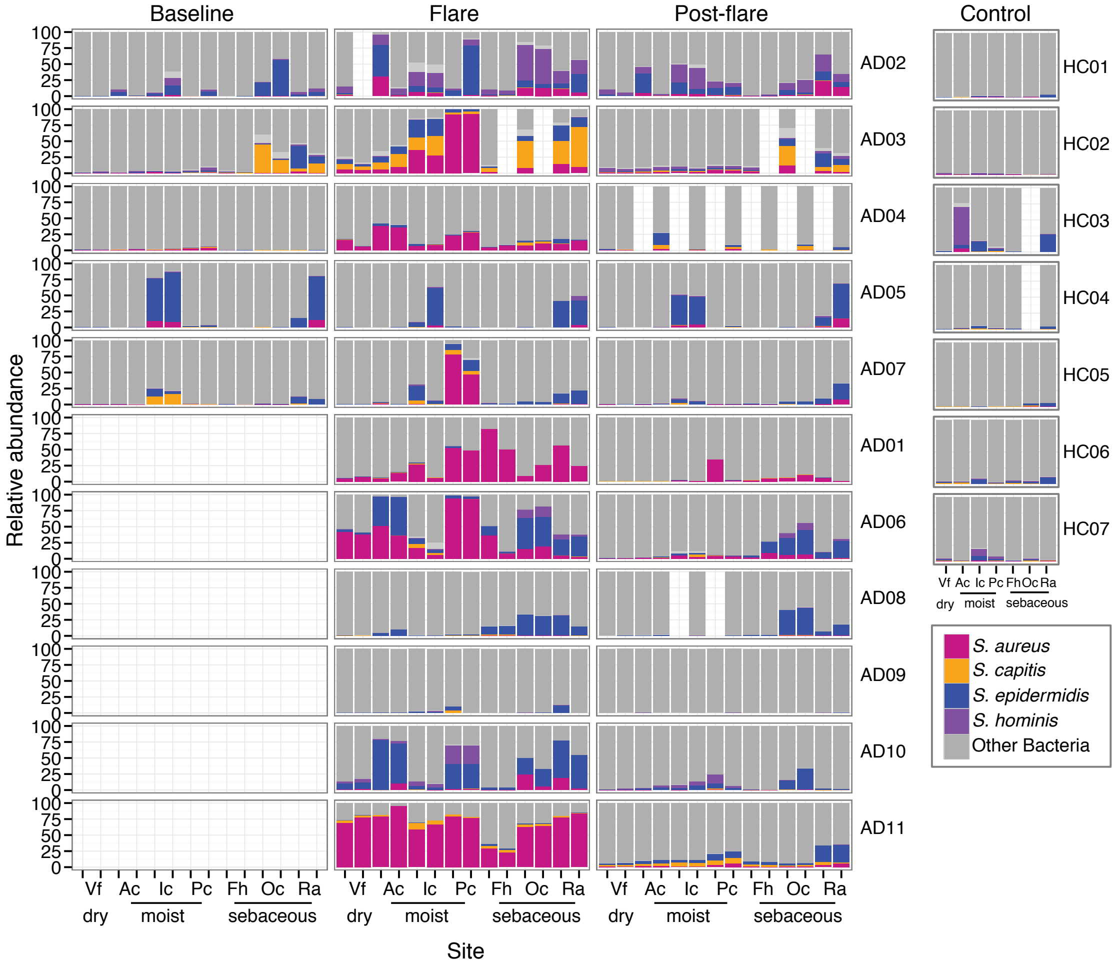


Figure 5.6. Relative abundance of staphylococcal species in relation to total bacterial population for all sites in AD patients and controls.

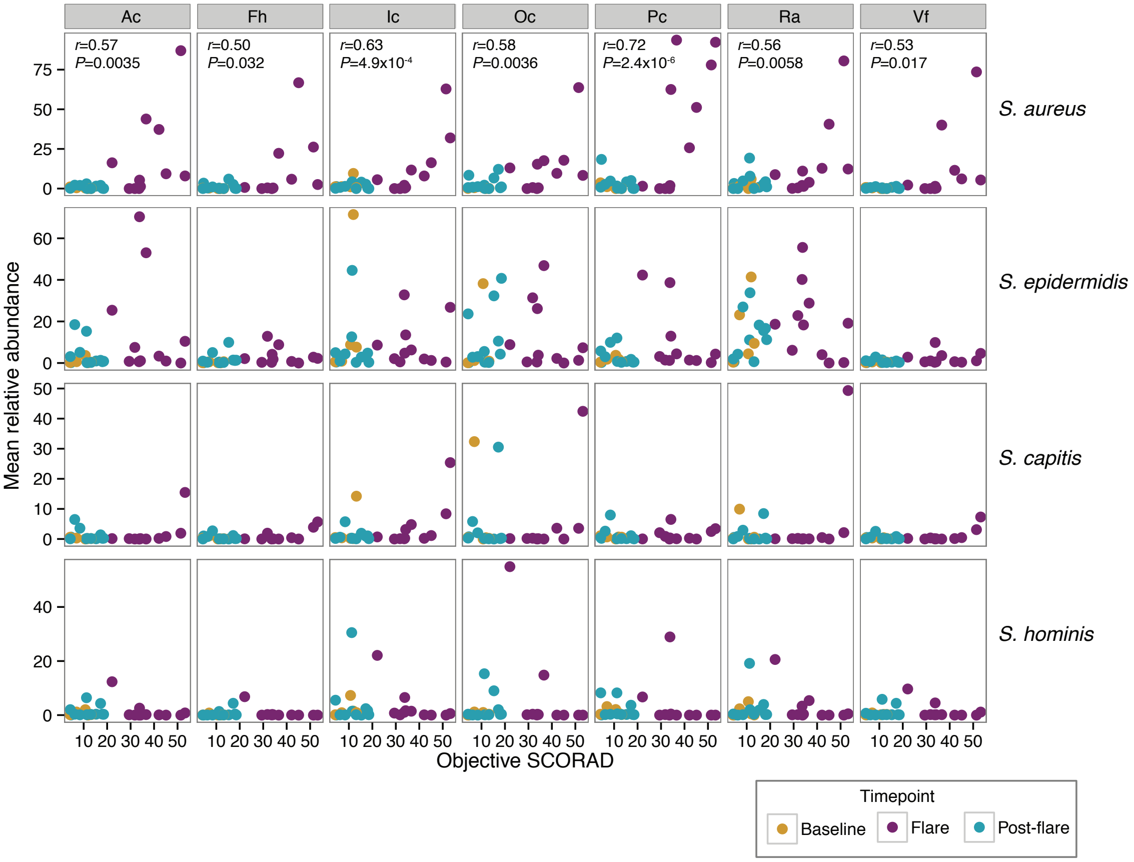


Figure 5.7. Correlation of various staphylococcal species mean relative abundance and Obj SCORAD for all sites of patients.

Partial correlation (adjusting for disease state). Only significant correlations are indicated.

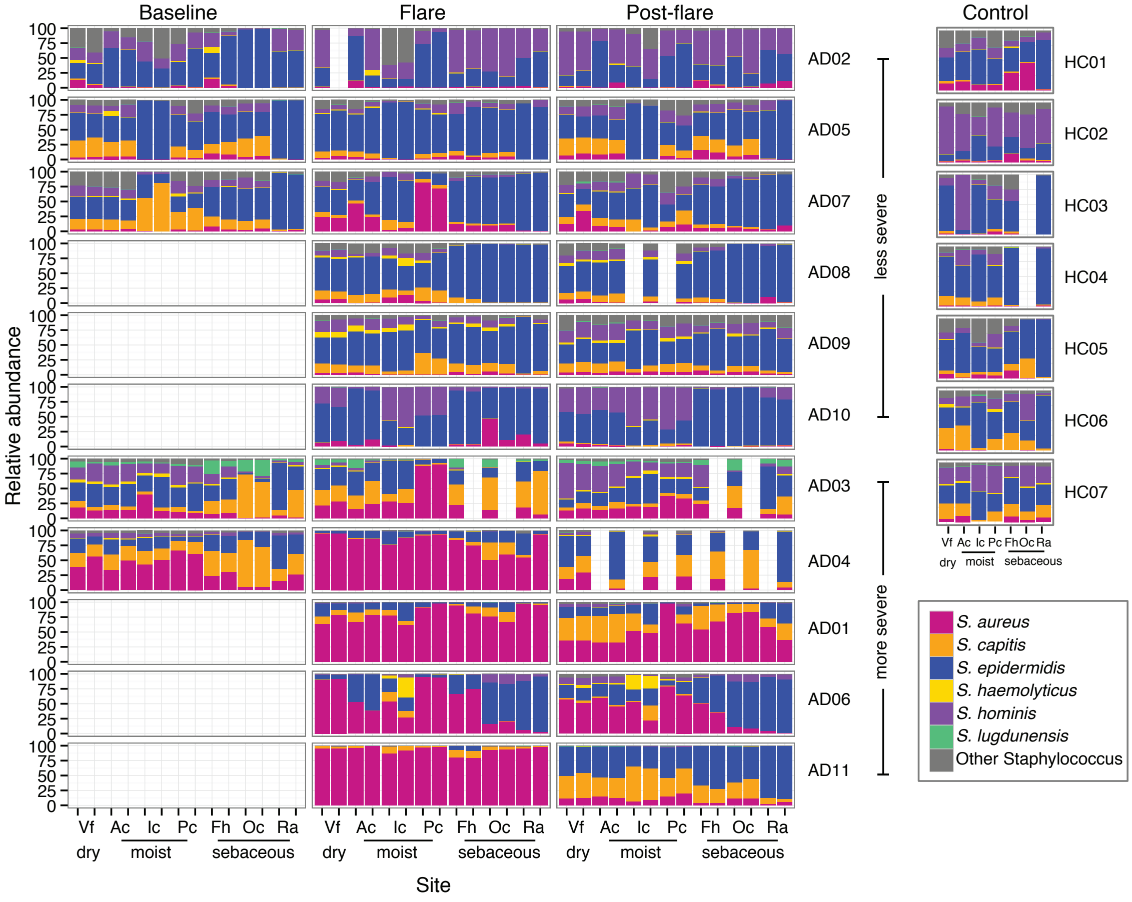


Figure 5.8. Relative abundance of staphylococcal species for all sites in AD patients and controls.

5.5 Distinct monoclonal *S. aureus* strains in more severe AD

Compared to traditional amplicon-based sequencing, shotgun metagenomics provides resolution of microbial communities at a strain and single nucleotide variant (SNV) level. Strain-level resolution is important as strains may exhibit functional differences. We used our previously validated strain-tracking approach to identify strains of *S. aureus* and *S. epidermidis* present in our AD patients (Oh et al., 2014; Oh et al., 2016). For S. *aureus* strain-tracking, microbial reads were mapped against a database composed of 215 *S. aureus* genomes, of which 61 representatives are shown in Figure 5.9A.

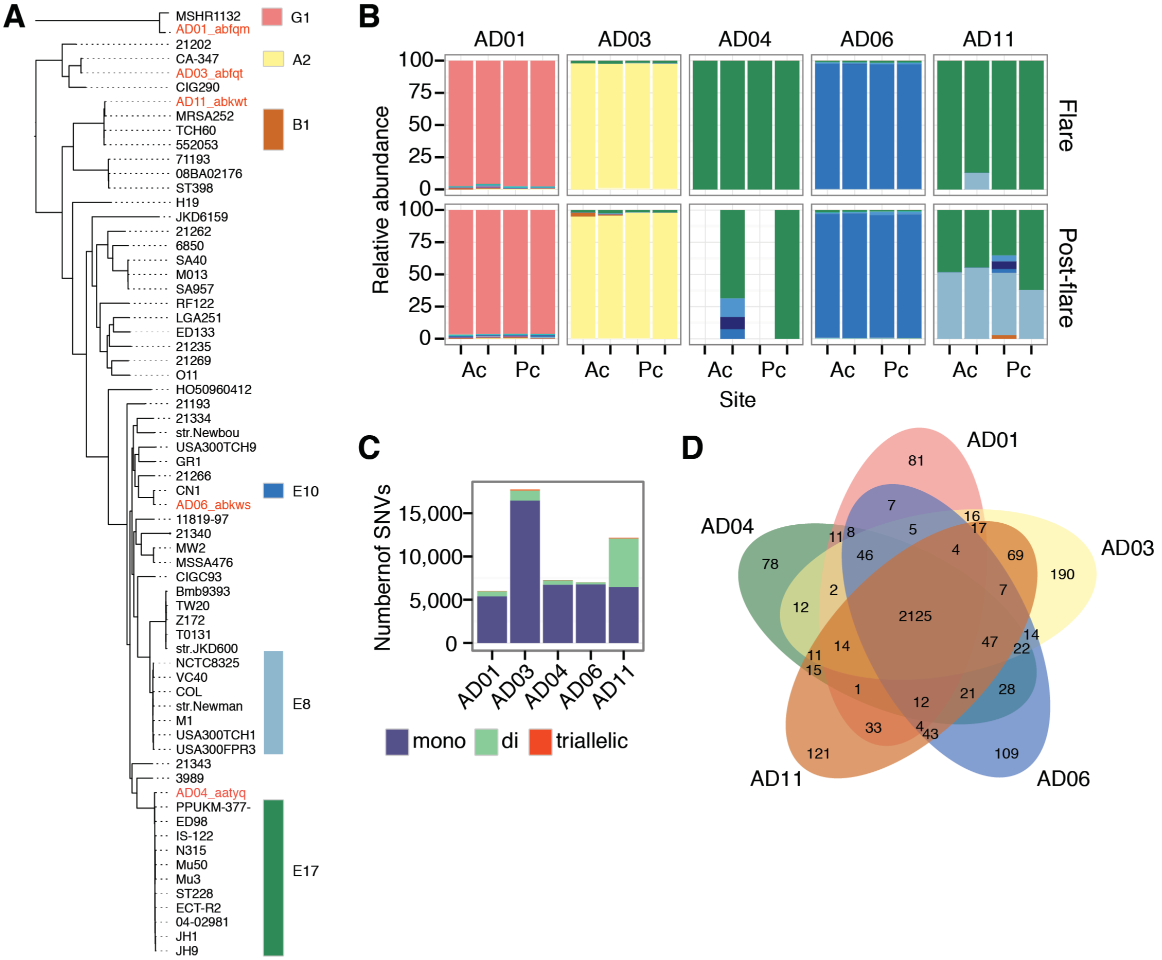


Figure 5.9. More severe AD patients are often monocolonized with a single *S. aureus* strain.

A) Dendogram of 61 representative *S. aureus* strains based on SNVs in the core genome. Strains labeled in red were isolated from patients in our study. Colors correspond to genomes of the same clade. Phylogenetically distant clade G1 is shown as an outgroup as it was recently reclassified as *S. argenteus (Tong et al., 2015b)*. B) For more severe AD patients, *S. aureus* clade relative abundances in bilateral antecubital (Ac) and popliteal creases (Pc) for AD disease states, flare and post-flare. Colors correspond to those in (A). C) For combined samples of all sites/time points of individuals in (B), barcharts show the number of SNVs per individual that are mono, di, and triallelic. D) Venn diagram showing the number of genes shared between isolates from patients in our study, indicated in red in (A).

In contrast to the heterogeneous communities of *P. acnes* and *S. epidermidis* strains observed in healthy adult skin (Oh et al., 2014; Oh et al., 2016), more severe AD patients were strikingly monocolonized with a single clade of *S. aureus* during disease flares [Figure 5.9B,

Figure 5.10]. For 4 out of the 5 severe AD patients, this monocolonization persisted in the post-flare. Patient AD11 was the exception, colonized by 3 different clades of *S. aureus* with clade E17 predominating during a flare and clades E8 and E17 predominating post-flare. The more severe AD patients were colonized with distinct *S. aureus* clades. This supports previous studies demonstrating AD patients did not share a single dominant *S. aureus* clone (Hoeger et al., 1992; Kim et al., 2009; Lomholt et al., 2005; Yeung et al., 2011). The variation in *S. aureus*-clades monocolonizing AD patients raises the possibility that this heterogeneity may contribute to the differential course and/or therapeutic responses of AD patients.

To confirm our strain-tracking results, we used a complementary approach in which SNVs were identified in the *S. aureus* core genome (1.9 Mbps shared between all sequenced *S. aureus*). To power this analysis, we combined all sites and time points for each patient. In total, we identified 38,867 variant positions in the *S. aureus* core or ~10,000 SNPs per patient. We then used the degree of polyallelism in each individual to infer genetic heterogeneity or the presence of multiple *S. aureus* strains. We calculated the number of mono, di, and triallelic SNVs for each patient [Figure 5.9C]. Consistent with strain tracking results, SNVs in *S. aureus*-monocolonized AD patients were diallelic at only 6.6% of sites, while heterogeneous patient AD11's SNVs were diallelic at 46% of sites.

S. aureus isolates cultured from each of the more severe AD patients underwent whole genome sequencing to confirm that the cultured patient isolates grouped into the respective clades predicted by strain-tracking of the metagenomic data [Figure 5.9A]. Based on standard cultivation methods and whole genome sequencing analysis, the five *S. aureus* isolates from the more severe AD patients were all methicillin-sensitive *S. aureus* (MSSA), consistent with higher incidences of MSSA than methicillin-resistant *S. aureus* (MRSA) cultivated from AD patient skin (Chaptini et al., 2015; Hsiang et al., 2012; Suh et al., 2008). Consistent with these patient isolates mapping to disparate phylogenetic clades, comparative genomic analysis revealed extensive heterogeneity in their gene content. The genome of a single *S. aureus* isolate encodes ~2,500 genes of which 67% (2125 genes) are present in every strain's genome and constitute the functional core [Figure 5.9D] while the remaining ~300 genes derive from the flexible pangenome comprised of 1,048 genes. Noncore genes showed functional enrichment in the KEGG pathways ko05150 *Staphylococcus aureus* infection and ko00906 Carotenoid biosynthesis. With a targeted search, enterotoxin genes, previously shown to exacerbate AD (Strange et al., 1996), were present in 4 of 5 AD patient strains of *S. aureus*. This strain-level gene variation generates additional questions regarding the potential role of specific strains on disease pathogenesis and host factors on clonal strain selection.

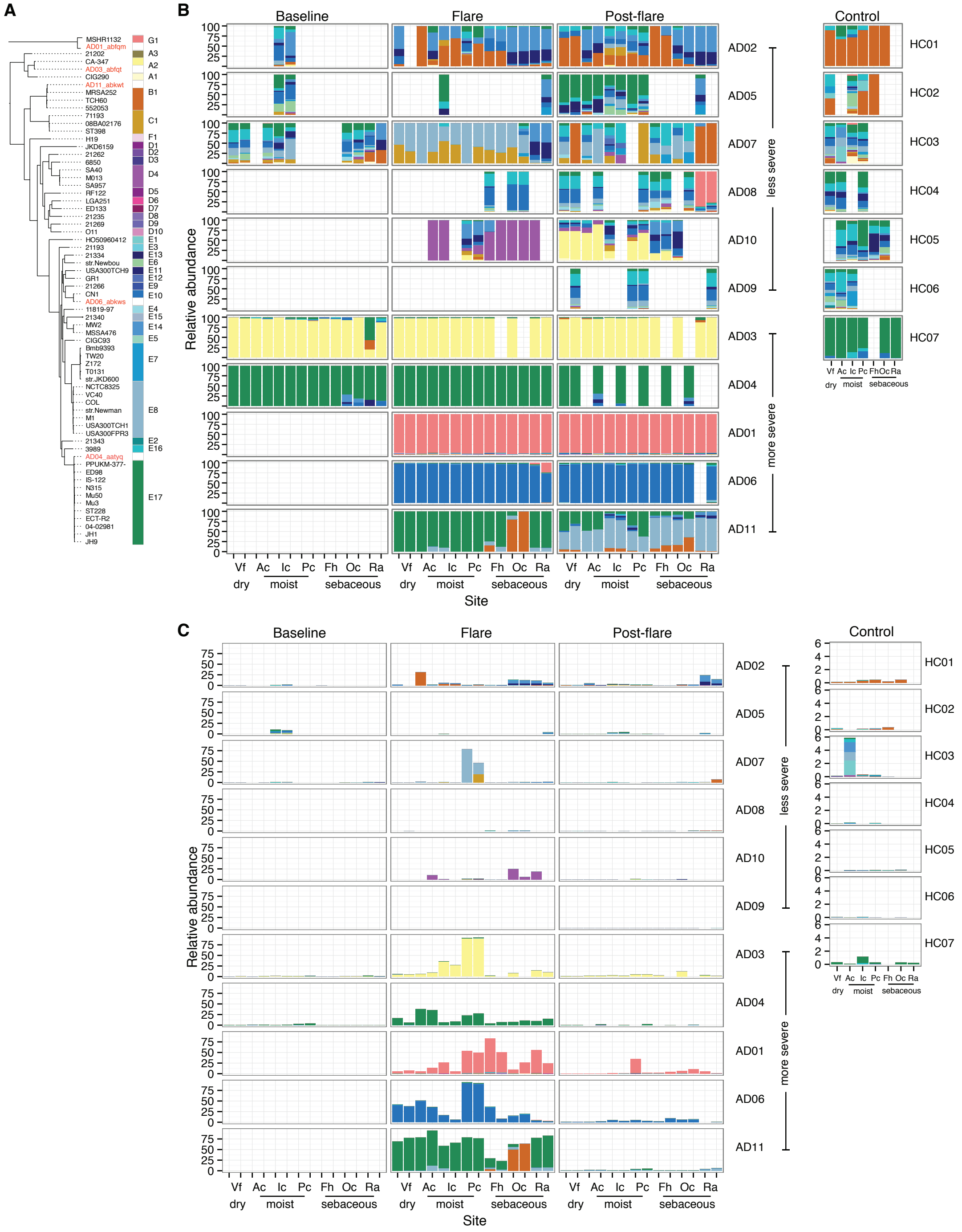


Figure 5.10. *S. aureus* clades for AD patients and controls.

(A) Cladogram of *S. aureus* strains based on SNVs in the core genome. Strains with names in red were isolated from patients in our study. Colors correspond to genomes of the same clade. Phylogenetically distant clade G1 is shown as an outgroup. (B) *S. aureus* clade relative abundances for all sites in AD patients and controls. Colors correspond to those in the (A). (C) *S. aureus* clade relative abundance normalized to percent *S. aureus* in the total bacterial population. Colors correspond to those in the (A).

5.6 Heterogeneous *S. epidermidis* strain communities

Next, we explored strain-level variation of the *S. epidermidis* isolates, using 61 sequenced *S. epidermidis* genomes [Figure 5.11A]. As seen with healthy adults (Oh et al., 2014) and children, AD patients' *S. epidermidis* communities at both flare and post-flare are composed of multiple different strains from diverse clades of the phylogenetic tree [Figure 5.11B and Figure 5.12], in direct contrast to the identification of homogeneous *S. aureus* communities. This multi-phyletic *S. epidermidis* strain diversity was observed for both the more severe and less severe AD patients [Figure 5.12]. However, analysis of the *S. epidermidis* strain composition in this cohort revealed a clustering of the less severe AD patients [Figure 5.11C]. Specifically, unsupervised clustering and principal coordinate analyses both identified *S. epidermidis* clades A29 and A30 as contributing to the clustering of the less severe patients and clade A20 as contributing to the clustering of the healthy adults [Figure 5.11D]. In contrast, the *S. epidermidis* strain diversity in healthy control children and more severe patients were intermixed.

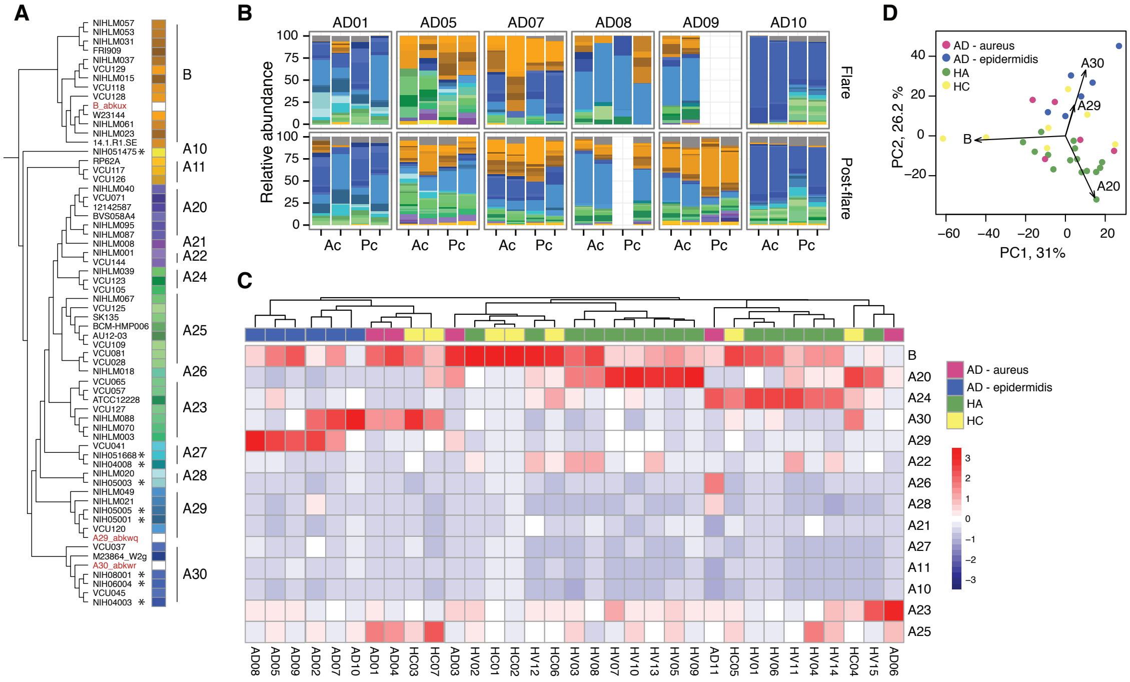


Figure 5.11. All patients are colonized by a multi-phyletic community of *S. epidermidis* strains.

A) Dendogram of *S. epidermidis* strains based on SNVs in the core genome. Red strains were isolated from patients in our study. Similar colors represent closely related strains that were grouped into 14 clades. Starred (*) isolates are nosocomial in origin B) For the less severe AD patients, *S. epidermidis* strain relative abundances in combined antecubital (Ac) and popliteal creases (Pc) for AD disease states, flare and post-flare. Colors correspond to those in (A). C) Heatmap shows mean relative abundance of each clade across all sites in *S. aureus* and *S. epidermidis*-predominant AD patients, healthy adults (HA), and healthy children (HC). D) In principal component analysis, clades A20, A29, and A30 drive separation between *S. epidermidis*-predominant AD patients and healthy adults.

In genomic analysis of the *S. epidermidis* clades, clades A29 and A30 were enriched in strains originally collected from nosocomial infections rather than as skin commensals (Conlan et al., 2012)(indicated with *s in Figure 5.11A). The overrepresentation of nosocomial isolates in AD patients suggests these strains may have the potential to outcompete commensals in inflammatory or non-steady-state conditions. Comparative genomic analysis of nosocomial isolates and the other strains revealed higher relative abundances of the SCC*mec* cassette (Conlan et al., 2012), which encodes genes necessary for methicillin-resistance, in the nosocomial isolates. To further evaluate the *S. epidermidis* strains in this cohort, isolates were collected from less severe patients AD05 and AD10. Whole genome sequencing identified the patient isolates as members of the A29 and A30 clade, respectively [Figure 5.11A in red]. Consistent with the trend of increased drug resistance genes observed through genomic analysis, these patient isolates were methicillin-resistant on cultivation.

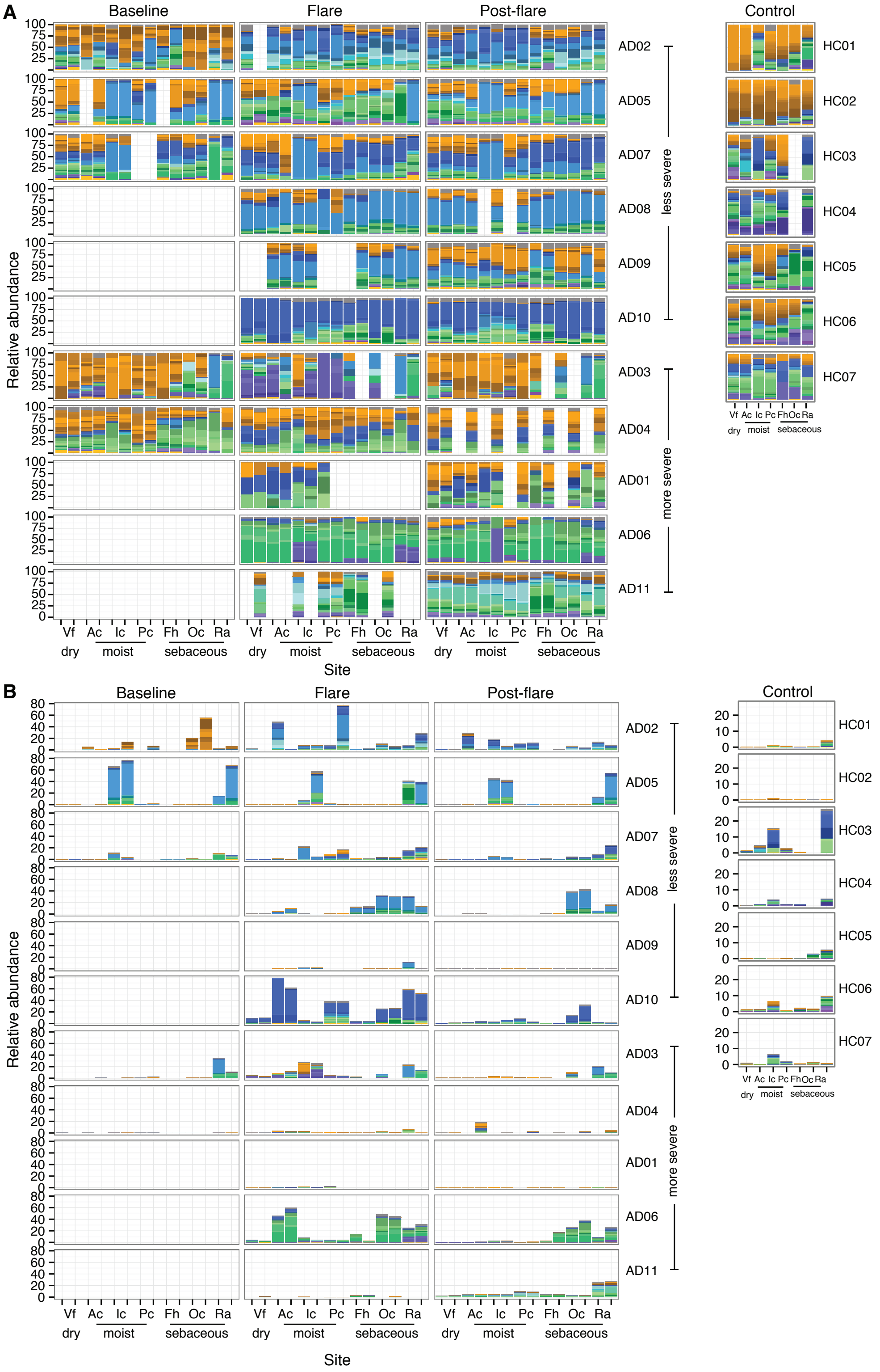


Figure 5.12. ***S. epidermidis*** **clades for AD patients and controls.**

A) *S. epidermidis* clade relative abundances for all sites in AD patients and controls. Colors correspond to those in Figure 4A. B) *S. epidermidis* clade relative abundance normalized to percent *S. epidermidis* in the total bacterial population. Colors correspond to those in **Figure 5.11**A.

5.7 Induction of AD-like immune responses in a murine model

While *S. aureus* has been tightly linked with AD, it remains unclear whether *S. aureus* can elicit and/or worsen AD skin disease or is a bystander that flourishes with increased access to extracellular matrix or other products of inflammation in eczematous skin (Cho et al., 2001; Kuusela, 1978). We next analyzed if AD patient strains would be sufficient to elicit skin inflammation in the absence of any known genetic predisposition or barrier disruption. To do this, we topically applied our clinically relevant AD patient staphylococcal strains onto intact skin of wild-type mice with a method previously developed to test the immune response to skin commensals [Figure 5.13A](Naik et al., 2015; Naik et al., 2012). We individually tested the five phylogenetically distinct *S. aureus* isolates from more severe AD patients and three *S. epidermidis* isolates (an isolate from clades A29, A30, and B). In contrast to the non-inflammatory responses observed following association with either skin commensals (Naik et al., 2015; Naik et al., 2012) or AD patient *S. epidermidis* isolates, topical application of the AD patient *S. aureus* isolates was sufficient to induce inflammatory responses as evidenced by epidermal thickening [Figure 5.13B,C, Figure 5.14A] as well as immune cell infiltrate composed of neutrophils and eosinophils [Figure 5.13D, Figure 5.14B].

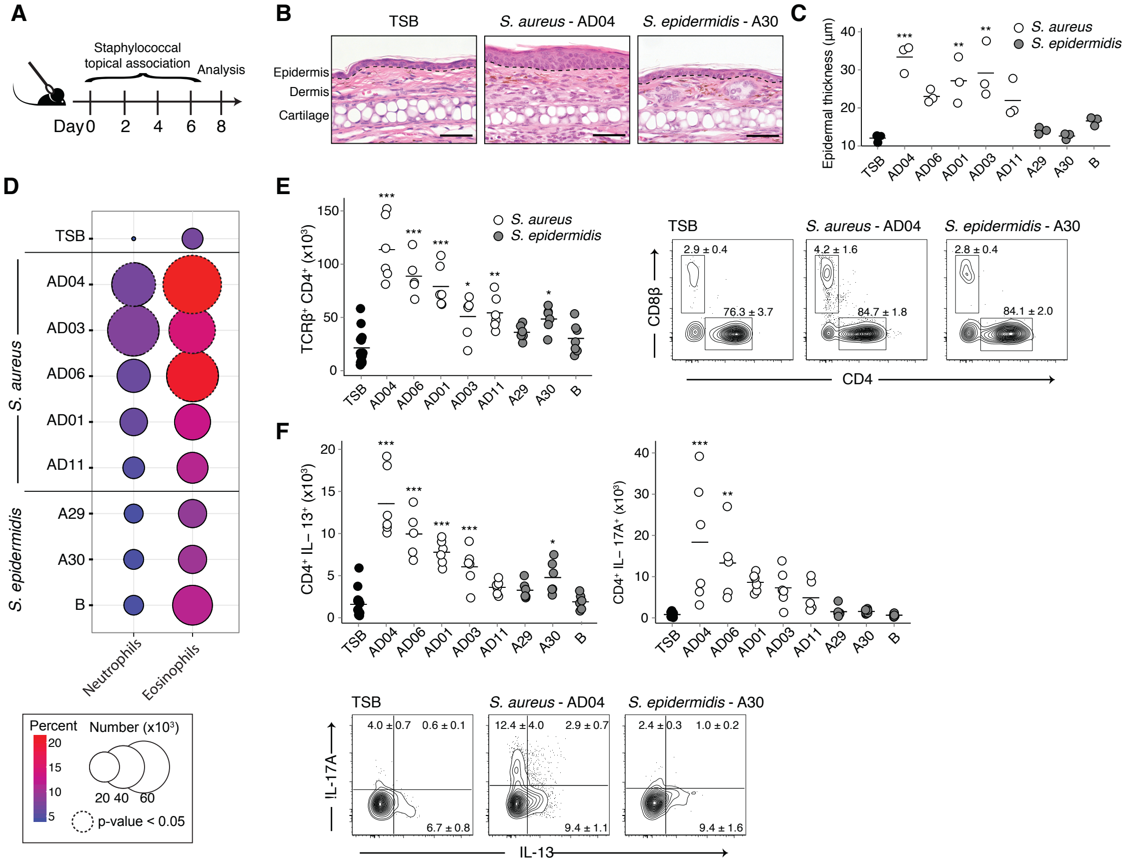


Figure 5.13. Topical application of AD isolates induce AD-like immune responses in murine models.

A) Mice were topically associated with staphylococcal monocultures every other day 4 times before sacrifice on the 8th day. B) Representative histological images of the ear pinnae of mice associated with tryptic soy broth TSB, *S. aureus* 2075, or *S. epidermidis* A30. Dotted line indicates separation between the epidermidis and dermis. Scale bar 50 µm. C) Epidermal thickness of ears post topical association of patient AD isolates. D) Of skin CD45$^+$ cells, the distribution of innate immune cells, neutrophils and eosinophils. Color corresponds to the percentage of the parent (Eosinophils: Lin$^-$, MHCII CD64$^-$; Neutrophils: Lin$^-$) population and size to the absolute number of cells. Isolates with significant differences compared to TSB (P-values < 0.05) have dashed outlines. E) Left: Absolute numbers of skin TCRβ$^+$ CD4$^+$ cells of mice in (D). Color indicates species designation of the isolate. Right: Flow plots show the frequencies of CD4$^+$ and CD8$^+$ effector T cells of mice in (B). F) Top: Absolute numbers of skin IL-13$^+$ and IL-17A$^+$ CD4$^+$ cells from mice in (D). Bottom: Frequencies of IL-13$^+$ and IL-17A$^+$ CD4$^+$ cells from mice in (D). Results are cumulative data from 2 independent experiments. *P<0.05, **P<0.01, ***P<0.001 as calculated by ANOVA with multiple comparison correction.

In addition, infiltration of T cell receptor (TCR) $\alpha\beta^+$ and $\gamma\delta^{low}$ cells was also observed [Figure 5.15A]. The degree of infiltration varied depending on the isolate, indicating strain-level specificity in ability to engage the immune system. More specifically, the dominant flare *S. aureus* isolates from AD04 and AD06 induced an amplified response compared to that of AD01, AD03, and the non-dominant strain from AD11. The majority of TCR-β^+ cells were CD4$^+$ [Figure 5.13E]. A proportion of these cells had the potential to produce the cytokine interleukin-13 (IL-13) [Figure 5.13F], an immune signature reminiscent of that seen in human AD patients. Cutaneous Th17 cells were also identified in *S. aureus* colonized mice [Figure 5.13F]. Recent reports have identified the presence of Th17 cells in AD lesions (Koga et al., 2008; Suarez-Farinas et al., 2013), particularly in certain patient populations (Noda et al., 2015). Similar to the CD4$^+$ T cells, the $\gamma\delta$ T cells of mice associated with *S. aureus* isolates also had the potential to make greater levels of interleukin-17A (IL-17A) [Figure 5.15B]. Overall, association of *S. aureus* strains isolated from AD patients to wild-type mice without barrier disruption induced AD-like immune responses in the skin. Thus, the findings suggest certain strains of *S. aureus* may be sufficient to exacerbate and/or elicit skin inflammation.

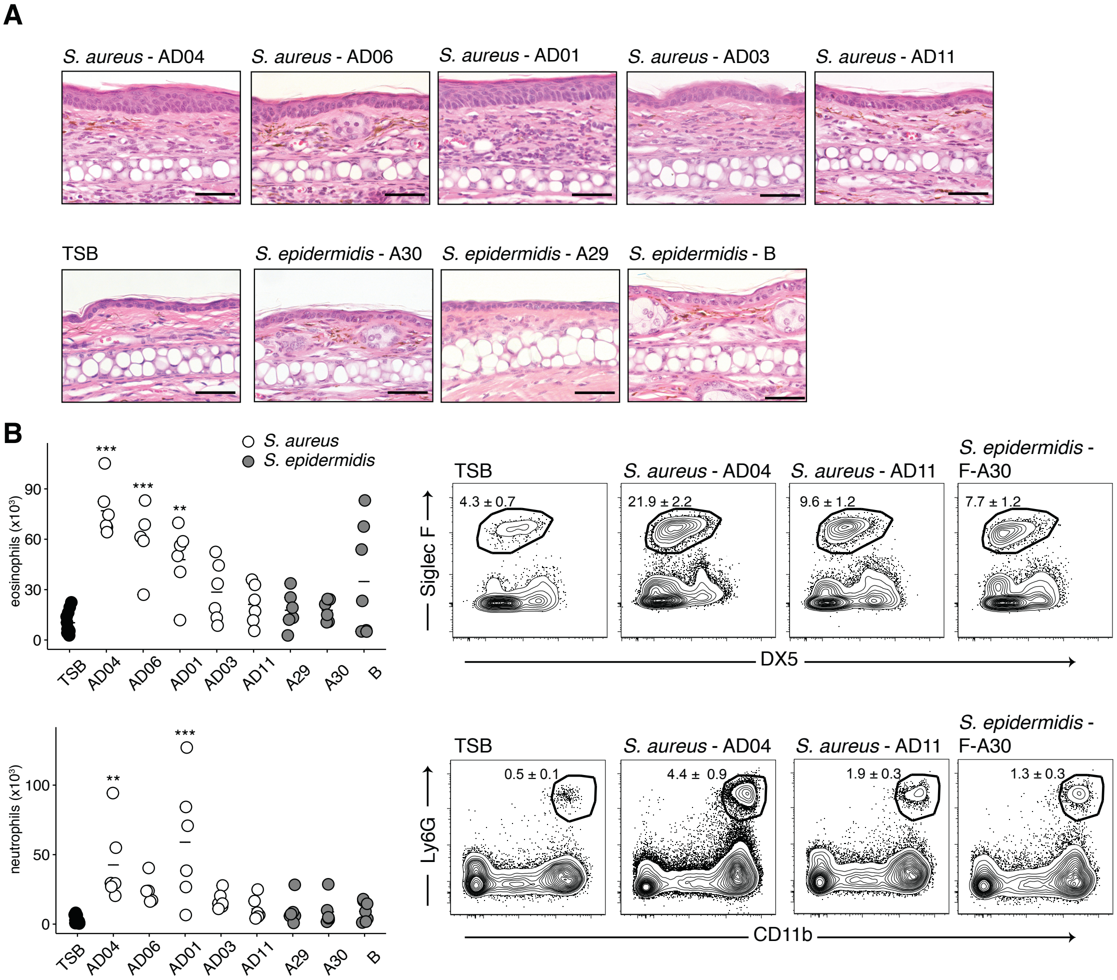


Figure 5.14. Histologic and cutaneous innate immune cell responses with AD isolate association in a murine model.

A) Representative histological images of the ear pinnae of mice associated with tryptic soy broth (TSB), and various AD patient *S. aureus* and *S. epidermidis* isolates. Scale bars, 50 µm. B) Absolute numbers and percentages of cutaneous eosinophils and neutrophils of mice in (A). *P<0.05, **P<0.01, ***P<0.001 as calculated by ANOVA with multiple comparison correction.

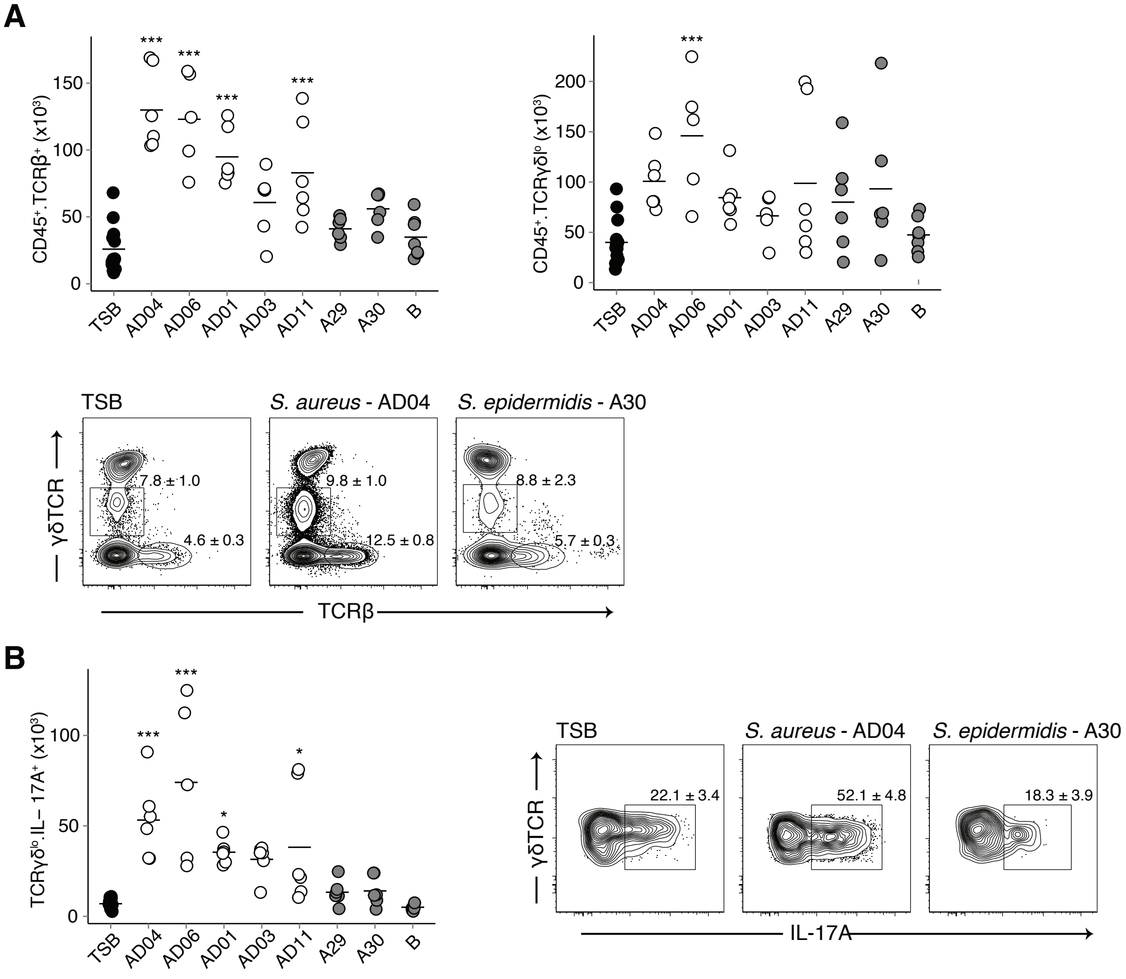


Figure 5.15. CD45$^+$ cutaneous immune responses with AD isolate association in a murine model.

A) Absolute numbers and representative flow plots of skin CD45$^+$ $\gamma\delta^{low}$ and TCRβ^+ cells with topical application of TSB and various AD patient *S. aureus* and *S. epidermidis* isolates. B) Absolute numbers and percentages of IL-17A$^+$ $\gamma\delta^{low}$ T cells. Results are representative of 2 independent experiments. *P<0.05, **P<0.01, ***P<0.001 as calculated by ANOVA with multiple comparison correction.

5.8 Discussion

AD is a heterogeneous disease with many contributing factors including skin barrier integrity, innate and adaptive immunity, and the microbiome. Here, we performed metagenomic sequencing to investigate the microbial communities of AD skin at strain and SNV-level resolution. Based on the skin microbiome, we stratified AD patients based on disease severity during flare. The differential abundances of *S. aureus* and *S. epidermidis*, along with the respective finding of MSSA and MRSE predominance may contribute to differential responses to therapies in AD patients (Bath-Hextall et al., 2010). Additional investigations of these microbiome phenotypic differences may improve the understanding of AD pathogenesis and lead to more targeted therapeutics. With strain tracking, we further stratified the more severe AD patients by the presence of unique *S. aureus* clades. Birth cohort studies may address whether these patients acquired bacterial strains from family members and/or environmental sources as part of microbial inheritance (Faith et al., 2015). Testing of *S. aureus* strains in gnotobiotic mice, similar to Bacteroides gut commensal studies, may functionally address whether monocolonization by *S. aureus* occurs through limited exposure or colonization resistance (Lee et al., 2013). Using strains isolated from inflamed AD skin, we examined the potential biological effects of variation between strains. With intact skin barrier and immunity, *S. aureus* was sufficient to induce AD-like features in a murine model, such as epidermal thickening and cutaneous infiltration of Th2 and Th17 cells. The magnitude of this effect varied depending on the isolated strain. Notably, colonization with AD11's non-dominant B clade *S. aureus* isolate, a strain similar to that identified in a pediatric control [Figure 5.10], induced minimal IL-13. In mouse models, *S. aureus* enterotoxins have been shown to act as superantigens that can initiate Th17 responses(Macias et al., 2011), while *S. aureus* δ-toxin can induce degranulation of mast cells (Nakamura et al., 2013). These genes were both present in the non-inducing *S. aureus* isolate indicating strain-variability exists not only in gene content but also expression of the genes. In the context of prior studies (Kobayashi et al., 2015; Naik et al., 2015; Naik et al., 2012; Nakamura et al., 2013), our findings demonstrate that AD-associated *S. aureus* strains can elicit skin inflammation in a manner distinct from other bacteria and suggests that cutaneous staphylococci may play an important role in AD skin inflammation.

In this study, we used metagenomic sequencing to examine strain-level microbial compositions of AD patient skin. We observed differential abundance of *S. aureus* and *S. epidermidis* associated with disease severity, suggesting potential differential phenotypes. Additionally, the less severe AD patients were colonized with more methicillin-resistant *S. epidermidis* strains while the more severe AD patients were colonized with methicillin-sensitive *S. aureus* strains. The more severe AD patients were also colonized by phylogenetically distinct, single strains of *S. aureus* during flares. With increasing recognition of highly individualized skin microbiomes (Oh et al., 2016), the different patient-specific strains underscore the individuality of the disease course and therapeutic response in AD patients and may represent an opportunity for precision medicine. The topical application of these patient-associated strains of *S. aureus* and *S. epidermidis* in a murine model demonstrated strain-specific differences in the ability to elicit AD-like histologic and immunologic features. Along with recent studies showing that early exposures can influence host immunity (Du Toit et al., 2015; Scharschmidt et al., 2015) and murine skin bacteria can exacerbate eczematous skin in an AD mouse model (Kobayashi et al., 2015), the current findings suggest that AD patient bacterial strains can induce AD-like inflammation in a host without skin barrier disruption or immune alterations. Thus, in light of the known links between severe AD and subsequent development of asthma and hay fever ("the atopic march"), targeted modulation of microbes that may play a role in pathogenesis of AD has the potential to abrogate development of atopic disorders.

5.9 Materials and Methods

5.9.1 Experimental design

Patients with AD and similarly aged healthy controls were recruited from the Washington DC metropolitan region, USA, between June 2012 and March 2015, to participate in a natural history study approved by the Institutional Review Board of National Human Genome Research Institute (http://www.clinicaltrials.gov/ct2/show/NCT00605878).

Eligibility criteria included age 2-18 years, moderate-to-severe disease, presence of ≥1 affected antecubital crease (inner elbow) or popliteal crease (behind the knee) at enrollment, and >3 weeks off of systemic antibiotics and corticosteroids. Patients were diagnosed with AD based on the UK Working Party definition (Williams et al., 1994a). Disease severity was measured by the objective SCORAD as assessed by one individual. At each clinical visit, objective SCORAD was used to determine study eligibility and disease status (Kunz et al., 1997; Oranje et al., 2007; Williams et al., 1994b). Moderate-to-severe disease was defined by objective SCORAD ≥ 15 (range 0-83) (Oranje et al., 2007).

For all subjects, exclusion criteria included receiving investigational new treatments, ultraviolet light therapy, monoclonal antibodies, systemic immunosuppressants within 7 days or five half-lives (taking the longer time period) of skin sampling, and clinically apparent underlying immunodeficiency. AD patients were also excluded if they took systemic antibiotics during the preceding three weeks (except for the post-flare timepoint). For healthy controls, additional exclusion criteria included current or prior chronic skin disease such as AD or psoriasis; asthma and allergic rhinitis, via International Study of Asthma and Allergies in Childhood questionnaire (Asher et al., 1995); other chronic medical conditions; and use of systemic antibiotics in the preceding 6 months.

Written consent was obtained from parents or guardians of all participating children. At all clinic visits, complete medical and medication history and skin examination was performed. To standardize skin sampling and optimize microbial load, no bathing, shampooing or emollients were permitted within 24 hours of sample collection. AD patients were sampled at three timepoints (baseline, flare, and post-flare) to capture the different stages of the chronic relapsing, remitting skin disease. Healthy controls were matched based on Tanner stage, which can be used to define an individual's stage of puberty based on physical examination. Unlike chronological age which does not necessarily correspond with a defined stage in sexual maturation, Tanner staging of sexual maturity can provide a phenotypic assessment of the physiologic age of an individual.

For AD patients in this study, baseline was defined as usual and stable disease state and ability to tolerate ≥7 days without topical AD treatments to intended sample sites and >2 weeks off both oral antibiotics and corticosteroids. The skin preparation regimen of 7 days without topical steroids or topical antimicrobial regimens prior to skin sampling was used to minimize the potential confounding effects of topical therapies on skin microbiota (Kong et al., 2012). Five of the 11 patients successfully reached a baseline state during the course of this study; the remaining patients required reinitiation of treatment due to clinical worsening of skin disease. Flare was defined as acute exacerbation of the disease on any skin site prior to initiation of intensified AD treatment and without restriction of usual treatments >24 hours prior to sampling. When skin disease worsening was apparent, patients were instructed to promptly contact the research team for evaluation and intensified skin-directed treatment. Post-flare was defined as 10-14 days after the initiation of intensified skin-directed AD treatment. Recommendations for intensified AD treatment included the following based on the patient's typical regimen: dilute bleach baths (0.25 cup of 6% bleach into bath half filled with water for a final concentrations of 0.0005%) two to four times per week, regular use of topical steroids twice daily, and bland emollients at least twice daily.

Seven sites were sampled bilaterally to represent the sites of disease predilection (antecubital creases and popliteal creases) and the different physiological characteristics of the skin [Figure 5.1]: dry (volar forearm/inner forearm), moist (antecubital crease, inguinal crease, popliteal crease), and sebaceous (glabella/central forehead, retroauricular crease/behind the ear, occiput/back of lower scalp). To obtain sufficient DNA for metagenomic sequencing, sites were sampled using a swab-scrape-swab procedure (Oh et al., 2014). All samples were stored in lysis buffer at -80C until DNA extraction. In addition to sequencing, swabs were taken from the antecubital crease, retroauricular crease, and the nares for culture analysis.

5.9.2 DNA extraction and sequencing of metagenomic samples

Procedures for library generation, sequencing, and processing of longitudinal samples were as described previously (Oh et al., 2014). Briefly, metagenomic DNA was prepared for sequencing using the Nextera DNA Library Prep Kit (Illumina) per manufacturer's instructions with the exception of increasing from 6 to 10 PCR cycles and increasing the AMPure XP Beads clean-up volume from 30uL to 50uL. Libraries were sequenced on an Illumina HiSeq at the NIH Intramural Sequencing Center to a target of 15 to 50 million clusters of 2 x 125bp reads. In total, for 18 individuals (11 patients and 7 controls), we obtained 422 samples or 2.26 trillon reads (191 Gb) of non-human, quality- filtered paired-end and singleton reads (median 2.4 million reads (.21 Gb) per sample). For sample processing, human reads were removed based on mapping to the hg19 + hg19 rRNA human reference genome, bases with quality scores below 20 were trimmed, and remaining reads less than 50bp were removed. To reduce computational burden, post quality control, samples with >20 million reads were subsampled to 10 million paired end reads, and singletons were discarded.

5.9.3 Taxonomic classification of skin species and diversity estimates

Microbial reads were assigned taxonomic classifications as previously described (Oh et al., 2014). Included in the microbial reference genome database are 2,342 bacteria, 389 fungal, 1,375 viral, and 67 archaeal genomes. In addition, a staphylococcus database was compiled from 315 complete and draft genomes from the National Center for Biological Information (NCBI, http://www.ncbi.nlm.nih.gov) as of October 2014. Nonhuman reads were separately mapped to both genome collections using bowtie2's –very-sensitive parameter with –k 10 to retrieve the top 10 hits (Langmead and Salzberg, 2012). The resulting alignment files were processed with Pathoscope v1.0 (Francis et al., 2013) to assign multiply mapped reads to their mostly likely genome of origin. Read hit counts were then normalized by genome and scaled to sum to one. Coverages of each output genome were calculated using genomeCoverageBed in the the Bedtools suite (Quinlan and Hall, 2010). To reduce the effects of spurious classifications from low abundance organisms, only species with ≥ 1 percent coverage of the genome were considered (Oh et al., 2014). For the multi-kingdom database, the Shannon diversity index was used for diversity comparisons. To reduce the potential for erroneous strain-calling, taxonomies were predicted at the species level by combing hits of strains within the same species.

5.9.4 Strain tracking of S. aureus and S. epidermidis

Strain tracking of the dominant flare species *Staphylococcus aureus* and *Staphyloccocus epidermidis* was performed as previously described (Oh et al., 2014). Briefly, reference databases for *S. aureus* and *S. epidermidis* were compiled from all complete and draft genomes available on NCBI, 215 and 61, respectively. For both species, whole genome alignment, with nucmer (Delcher et al., 2002), was then used to identify the "core" region shared between all sequenced strains. SNVs identified in these core regions were subsequently used to generate dendograms with PhyML 3.0. Based on the dendograms, we grouped strains into subtypes or clades, 34 for *S. aureus* and 14 for *S. epidermidis.* Due to redundancy in many of the *S. aureus* draft genomes, a consolidated tree composed of all 42 complete genomes and 19 representative draft genomes was generated for visualization purposes [Figure 5.10A]. For strain-tracking to avoid noise from other staphylococcal species, metagenomic reads were first filtered against the staphylococcus database minus the species being strain-tracked (--very-sensitive, -score-min L,- 0.6,0.006). The remaining reads were then mapped to each species database with bowtie2 (--very-sensitive, -score-min L,- 0.6,0.006, -k number of genomes)(Langmead and Salzberg, 2012) with zero tolerance for mismatches. The resulting alignment file was then processed with Pathoscope (-theta_prior 10 x 10^88) (Francis et al., 2013) to deconvolute multiple mapping reads. Accuracy of this strain-tracking approach was previously validated with extensive simulations (Noda et al., 2015).

5.9.5 Identification of SNVs in the S. aureus core

To achieve sufficient coverage for SNV analysis, all samples for each *S. aureus*-predominant patient were combined. For SNV analysis as described previously (Oh et al., 2016), metagenomic reads were mapped against the *S.aureus* core genome using bowtie2 (--very-sensitive). The resulting alignment file was sorted by samtools and then processed with GATK's IndelRealigner (McKenna et al., 2010). The corrected alignment file was analyzed with samtools and bcftools to identify possible variants (samtools mpileup -uD -q30 -Q30, bcftools view -Abvcg, vcfutils.pl varFilter -D99percentileofcoverage -d4 -1 .00001 -4 .00001). Custom scripts were then used to filter possible variants based on criteria described in (Lieberman et al., 2014). Briefly, an alternate allele was only considered if it was supported by >2 reads with a minimum mapping quality of 30, had an allele frequency >3%, and fewer than 20% of reads supporting the SNV also mapped to an indel. Due to limited numbers of reads in some of the patients, SNVs were detected in subsamples of 350,000 reads for each patient.

5.9.6 Patient isolate collection, genome sequencing, and annotation

Skin and nasal cultures were obtained with Catch-all Collection Swabs (Epicentre) pre-moistened with Fastidious Broth (Remel), placed in 2.0ml Fastidious Broth supplemented with 10% glycerol, and frozen at -80°C. Swabs were thawed, vortexed, serial diluted, and plated on Tryptic Soy Agar with 5% Sheep Blood (Remel). After overnight incubation at 37°C, colonies were picked and stored in LB with 20% glycerol. Colonies were screened by PCR for *S. aureus* using Nuc1 (5′-GCGATTGATGGTGATACGGTT-3′) and Nuc 2 (5′ AGCCAAGCCTTGACGAACTAAAGC-3′), or *S. epidermidis* using Se705-1 (5’-ATCAAAAAGTTGGCGAACCTTTTCA-3’) and Se705-2 (5’-CAAAAGAGCGTGGAGAAAAGTATCA-3’) as previously described (Martineau et al., 1996; Zhang et al., 2004). Individual colonies were then streaked on blood agar for two passages. Isolates were grown overnight in Tryptic Soy Broth at 37C, pelleted with centrifugation, and genomic DNA was extracted using the Promega Maxwell Tissue

DNA Kit with the addition of Readylyse Lysozyme Solution (Epicentre) and Lysostaphin (Sigma). DNA was treated with RNase, re-purified with the Genomic DNA Clean and Concentrator Kit (Zymo), and quantified using a Nanodrop spectrophotometer and Qubit (ThermoFisher). 1.0ng of bacterial DNA was used as input into the Nextera XT Sample prep kit (Illumina) as suggested by manufacturer.

Nextera libraries were generated from the genomic DNA and sequenced using a paired-end 300-base dual index run on an Illumina MiSeq to generate 1 million to 2 million read pairs per library for ~80x genome coverage. Reads for each isolate were assembled with MaSuRCA (version 2.2.1) (Zimin et al., 2013) or SPAdes (version 3.6.0) (Bankevich et al., 2012). Best k-mer length estimates on paired-end reads were evaluated using KmerGenie (version 1.6300)(Chikhi and Medvedev, 2014) and utilized in running the MaSuRCA assembler for each genome. The SPAdes assembler was run using K-mer values of 21, 33, 55, 77, 99, and 127. Contigs >= 500 nt were retained. For comparative genomic analysis, genome annotation was done using the GS Analysis Engine (http://ae.igs.umaryland.edu/cgi/index.cgi, (Galens et al., 2011)). For upload to NCBI, genome annotation was done using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP: http://www.ncbi.nlm.nih.gov/genome/annotation_prok/). Distribution of genes between the 5 *S. aureus* assemblies was visualized with jvenn (Bardou et al., 2014).

5.9.7 Methicillin Resistance

Glycerol stocks from clinical isolates as well as control strains (ATCC:BA1556; ATCC:29213) were plated overnight on blood agar. Individual colonies were then plated on Mannitol Salt Agar (Remel) and Mannitol Salt Agar with Oxacillin (Remel). Plates were scored for Mannitol fermentation and growth or no growth at 35°C for 24hrs.

To verify these results with a Cefoxitin Disk assay, individual colonies were picked and completely resuspended in SOC Broth. The suspension was plated on Mueller-Hinton Agar (Remel) and allowed to dry 5 minutes. One 30 µg Cefoxitin Disk (BD) was placed on the plate and incubated at 35°C for 24hr. Zones of inhibition were measured and scored as described (CLSI, 2013). Briefly, *S. aureus* was scored as susceptible ≥22mm and resistant ≤21mm. *S. epidermidis* was scored as susceptible ≥25mm and resistant ≤24mm.

5.9.8 Mice

C57BL/6 specific pathogen free (SPF) mice were purchased from Taconic Farms. All mice were bred and maintained under pathogen-free conditions at an American Association for the Accreditation of Laboratory Animal Care (AAALAC)-accredited animal facility at the NHGRI and housed in accordance with the procedures outlined in the Guide for the Care and Use of Laboratory Animals. All experiments were performed at the NHGRI under an animal study proposal approved by the NHGRI Animal Care and Use Committee. Female mice between 6 and 12 weeks of age were used for each experiment. In general, each mouse of the different experimental groups is reported. Exclusion criteria such as inadequate staining or low cell yield due to technical problems were pre-determined. Animals were assigned randomly to experimental groups.

5.9.9 Topical association

Topical association of mice was based on (Naik et al., 2015). *S. aureus* strains AD01_abfqm, AD03_abfqt, AD04_aatyq, AD06_abkws, AD11_abkwt and *S. epidermidis* strains A29_abkwq, A30_abkwr, and B_abkux, isolated from AD patients, were cultured in tryptic soy broth at 37°C for 18h. Before topical application, bacteria were enumerated by assessing colony-forming units using traditional bacteriology techniques and by measuring optical density (OD) at 600 nm using a spectrophotometer.

For topical association, a sterile epicenter Catch-All swab was moistened in liquid culture of the bacteria and then rubbed against the ears of mice until they became visually moist. Topical association was repeated every other day four times. For each experiment,18h cultures were normalized using OD600 to achieve similar bacterial density (approximately 10^8 c.f.u. per ml). Mice were euthanized 8 days after the first topical association with bacteria.

5.9.10 Tissue processing

Cells from the ear pinnae of mice were isolated as previously described (Naik et al., 2012). Briefly, ears were excised and separated into dorsal and ventral sheets. Tissue samples were digested in RPMI 1640 containing 2 mM L-glutamine, 1 mM sodium pyruvate and nonessential amino acids, 20 mM HEPES, 100 U/ml penicillin, 100 mg/ml streptomycin, 50 mM β-mercaptoethanol, and 0.25 mg/ml Liberase purified enzyme blend (Roche Diagnostic Corp.) and incubated for 1 hour 45 minutes at 37°C in 5% CO2. Digested skin sheets were homogenized using the Medicon/Medimachine tissue homogenizer system (Becton Dickinson).

5.9.11 In vitro restimulation

For detection of basal cytokine potential, single-cell suspensions from ear tissue were cultured directly ex vivo in a 96-well U-bottom plate in complete medium (RPMI 1640 supplemented with 10% fetal bovine serum (FBS), 2 mM L-glutamine, 1 mM sodium pyruvate and nonessential amino acids, 20 mM HEPES, 100 U/ml penicillin, 100 mg/ml streptomycin, 50 mM β-mercaptoethanol) and stimulated with 50 ng/ml phorbol myristate acetate (PMA) (Sigma-Aldrich) and 5 mg/ml (mouse) ionomycin (Sigma-Aldrich) in the presence of brefeldin A (GolgiPlug, BD Biosciences) for 2.5 h at 37°C in 5% CO2. After stimulation, cells were assessed for intracellular cytokine production as described below.

5.9.12 Flow cytometric analysis

Murine single-cell suspensions were incubated with fluorochrome-conjugated antibodies against surface markers CD4 (clone RM4-5), CD8β (eBioH35-17.2), CD11b (M1/70), CD11c (N418 or HL3), CD19 (6D5), CD45.2 (104), CD49b (DX5), CD64 (X54-5/7.1), Ly6G (1A8), MHCII (M5/114.15.2), NK1.1 (PK136), TCRγδ (GL3), TCRβ (H57-597), and/or SiglecF (E50-2440) in Hank's buffered salt solution (HBSS) for 20 min at 4°C and then washed. LIVE/DEAD Fixable Blue Dead Cell Stain Kit (Invitrogen Life Technologies) was used to exclude dead cells. Cells were then fixed for 30 min at 4°C using the eBioscience fixation kit and washed twice with corresponding permeabilization buffer. For simultaneous Foxp3 and intracellular cytokine staining, cells were stained with fluorochrome-conjugated antibodies against Foxp3 (FJK-16 s), IFN-γ (XMG- 1.2), IL-13 (eBio-13A) and IL-17A (eBio17B7) in permeabilization buffer (eBioscience) for 1 hr at 4°C. Each staining was performed in the presence of purified anti-mouse CD16/32 (93), 0.2 mg/ml purified rat IgG and 1 mg/ml of normal mouse serum (Jackson Immunoresearch). All antibodies were purchased from eBioscience, Biolegend, or BD Biosciences. Cell acquisition was performed on a Fortessa flow cytometer using FACSDiVa software (BD Biosciences) and data were analyzed using FlowJo software (TreeStar).

5.9.13 Histology

Mice were euthanized on day 8 after topical application of the AD patient isolates. TSB associated mice were used as controls. The ears from each mouse were removed and fixed in PBS containing 10% formalin. Paraffin-embedded sections were cut at 0.5 mm, stained with haematoxylin and eosin and examined histologically.

5.9.14 Statistics

All statistical analyses were performed in R and the majority of graphs generated with ggplot2 (Wickham, 2009). Data are represented as mean ± standard error of the mean unless otherwise indicated. As disease severity differed minimally from left to right symmetric sites, left and right values were averaged in relative abundance plots and prior to statistical comparisons. AcPc indicates the mean of values from samples of the antecubital and popliteal crease for each individual (post-averaging of left and right symmetric sites). To avoid repeated measures when all sites were considered, samples belonging to an individual were averaged before statistical comparisons between timepoints when using summary metrics such as means, diversity, or theta indices.

Pearson correlations of non-zero values were used for all partial correlations adjusting for disease state (pcor.test in R package ppcor). For all boxplots, center lines represent the median and edges the first and third quartiles. The nonparametric Wilcoxon rank-sum test was used to determine statistically significant differences between populations (wilcox.test in R). Where indicated, within-subject analysis was performed with option "paired=T" in wilcox.test. All *P*-values were adjusted using p.adjust in R using Bonferroni (# comparisons ≤ 10) or false discovery rate (# comparisons > 10) corrections. Statistical significance was ascribed to an alpha level of the adjusted *P*-values ≤ 0.05. Similarity between samples was assessed using the Yue–Clayton theta, which assesses the similarity between two samples based on (1) number of features in common between two samples, and (2) their relative abundances with $\theta = 0$ indicating totally dissimilar communities and $\theta = 1$ identical communities (Yue and Clayton, 2005).

For functional experiments, mice were assigned randomly to groups. Mouse studies were not performed in a blinded fashion. Generally, each mouse of the different experimental groups is reported. Statistical significance was determined by ANOVA with multiple comparison correction (aov and TukeyHSD in R).

CHAPTER 6 Moving forward: From description to function

In summary, this thesis provides an analysis of the skin microbiome in health and disease at previously unexplored resolution. Analysis at this level was possible due to technical advances in DNA extraction techniques optimized for the microbially diverse, yet low biomass, environment of the skin and the creation of novel software pipelines that exploited the depth of information available in whole genome metagenomic sequencing (Chapters 2 and 3). With strain-tracking tools, it was discovered that the stability of healthy adult skin microbial communities is driven by the persistence of heterogeneous communities of *P. acnes* and *S. epidermidis* strains (Chapters 3 and 4). In patients with atopic dermatitis (Chapter 5), these analysis techniques showed that all patients harbor heterogeneous communities of *S. epidermidis*, like their healthy counterparts. Interestingly, a subset of those patients was also colonized with a single clade of *S. aureus* across body sites. Functional studies revealed these *S. aureus* isolates were sufficient to induce skin inflammation in mice. Overall these results provide comprehensive descriptions of the microbial communities present on the skin in steady state and inflammatory conditions. Moving forward many questions remain regarding function, what role are the microbes playing in the skin in different contexts; and ecology, how did these communities assemble?

6.1 Function

DNA sequencing is a useful, unbiased, tool for revealing the microbes present in a sample. However, it has the limitation of not discriminating between live colonized microbes and dead transients. Traditional culture techniques can differentiate the two;

with the caveat that culture conditions can wildly skew the results. RNA sequencing would address the issue by revealing microbes' functional activity. In fact, a metatranscriptomic analysis of follicular contents has revealed the microbes behave differently in healthy individuals and acne patients (Kang et al., 2015). Unfortunately, due to the low biomass of the skin, obtaining sufficient microbial RNA for analysis is a nontrivial process and even more efficient extraction techniques are necessary. However, as a proxy for a microbe's activity, Korem *et al.* developed an analysis technique that compares read distributions at the origin of replication and elsewhere in the genome as evidence of active bacterial replication (Korem et al., 2015). Understanding which microbes are active in health and disease will help dedicate which microbes to target in the development of therapies.

6.2 Ecology

Results from strain-level analyzes allow many hypotheses to be generated as to how the communities assembled. For example, one could speculate that *S. epidermidis* strains exist as heterogeneous communities because they act cooperatively together, while *S. aureus* strains exist alone as they actively inhibit one another. With appropriately large samples sizes, statistical power would be sufficient to predict competition or cooperation between strains. Detected interactions could then be tested with competition assays measuring interactions between strains or species. To increase accuracy, isolates matching those in the metagenomic sequences should be tested first. Then, further testing with additional isolates could show how conserved the relationship is. Once interactions are confirmed, further in depth experiments would be necessary to resolve the mechanism. Understanding microbial relationships will be helpful in determining whether live microbes or microbial products could eventually be employed to promote optimal skin health.

In total, the microbial communities reported in this thesis have been described in greater detail than ever before. Moving forward, I hope this data serves as the foundation on which many more testable hypotheses are generated and that the AD story will be continued such that future mechanistic animal studies will provide information that can ultimately lead to the development of targeted therapeutics for use in patients [Figure 6.1].

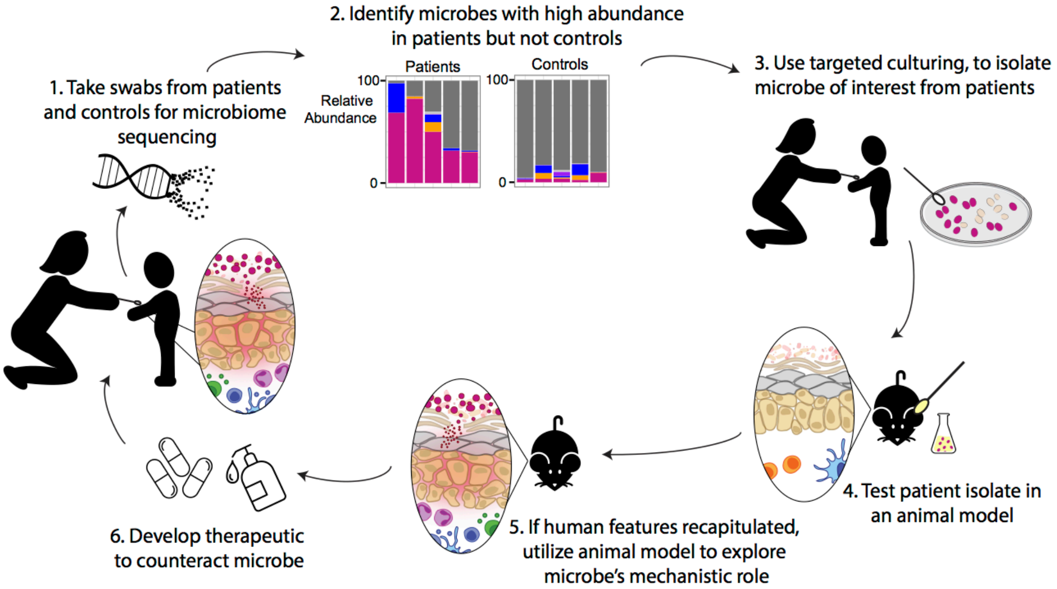

2. Identify microbes with high abundance in patients but not controls
Patients
Controls
100
Relative Abundance
0
100
0
1. Take swabs from patients and controls for microbiome sequencing
3. Use targeted culturing, to isolate microbe of interest from patients
4. Test patient isolate in an animal model
5. If human features recapitulated, utilize animal model to explore microbe's mechanistic role
6. Develop therapeutic to counteract microbe

Figure 6.1. : Formulating testable hypotheses from sequencing data to generate novel therapeutics.

The following diagram shows how microbial sequencing data from healthy controls and patients can be used to generate hypotheses about putative causative microbes. Computationally identified microbes of interest can then be isolated from patient swabs utilizing targeted culturing methods. These microbes can then be tested in animal models to decipher the microbe's possible mechanistic role in a disease. These results can then be utilized to develop a therapeutic to counteract the microbe.

CHAPTER 7 APPENDIX: Adapting Koch's postulates

7.1 Introduction

In the late 19th century, Robert Koch established his famous postulates as stringent guidelines to evaluate causation in infectious disease (Koch, 1890). These original postulates require isolation of the putative pathogen and reinfection of a healthy host to prove causation. Over the years, Koch's postulates have been continually restated to incorporate the latest scientific findings and technologies (Evans, 1976; Falkow, 1988; Fredricks and Relman, 1996; Rivers, 1937). Modern molecular techniques have demonstrated that current or previous members of a microbial community can affect disease outcome, providing a nuanced view of strict causation as originally proposed by Koch. There is thus a need to incorporate microbial communities into rigorous modern guidelines for evaluating disease causation.

Note: The work presented in this appendix has been previously published in (Byrd and Segre, *Science* 2016).

7.2 1 Pathogen = 1 Disease

Koch's original postulates can be summarized as follows: First, the microorganism occurs in every case of the disease; second, it is not found in healthy organisms; and third, after the microorganism has been isolated from a diseased organism and propagated in pure culture, the proposed pathogen can induce disease anew. Koch did not include the often cited fourth postulate that the microorganism must then be reisolated from the experimentally infected host, but it has come to be viewed as necessary to complete the loop asserting causation. Although revolutionary for the time, the postulates have since been a double-edged sword. For example, the third postulate was implemented to guard against misassignment of causality due to mixed cultures. However, blind adherence to this postulate would mean excluding obligate parasites and viruses as infectious agents.

Over the past decade, sequencing technologies and advanced analytic tools have enabled whole-genome sequencing of both microbial isolates and communities. These advances raise new questions of how Koch's postulates can be updated to incorporate these molecular techniques. For example, when assigning causality to an organism, can a fully sequenced genome act as a surrogate for pure culture, even when the suspected organism requires additional microbes for successful propagation? Also, how do you address the role of microbial communities in disease pathogenesis? Here we address these questions and introduce new variables into Koch's one organism = one disease equation.

7.3 1 Pathogen + 1 Colonization resistor = 0 Disease

A modern test of Koch's postulates is the risk to patients of becoming colonized with a pathogen while hospitalized. In this setting, it has become clear that some commensal organisms can protect the host against pathogenic enemies, a process termed “colonization resistance.” These commensal protectors defend the host either by directly inhibiting the pathogen or by enhancing host immunity (Buffie and Pamer, 2013). Recent evidence for both varieties of colonization resistance highlights how the presence of specific commensal bacteria can alter the pathology induced by Koch-verified infectious bacteria (Buffie et al., 2015; Schieber et al., 2015). These studies demonstrate how microbial community sequencing can be used to differentiate when an infectious agent induces disease in some but not all hosts.

In one study, Buffie *et al.* (Buffie et al., 2015) showed that mice treated with antibiotics exhibited varied susceptibility to infection by *Clostridium difficile*, a major cause of antibiotic-induced diarrhea. The authors also performed a similar analysis with a cohort of patients undergoing stem-cell transplant. Because of antibiotic treatment and compromised immune function, these patients are particularly susceptible to *C. difficile* infection. With microbial community sequencing and subsequent modeling of microbial interactions, the authors identified *Clostridium scindens* as a commensal associated with colonization resistance. This was validated when mice precolonized with a commercially available strain of *C. scindens* exhibited amelioration of symptoms associated with *C. difficile* infection. Mechanistically, it was demonstrated that *C. scindens* modifies endogenous bile acids to inhibit *C. difficile* growth (Buffie et al., 2015; Sorg and Sonenshein, 2010). Buffie *et al.* provide a well-validated example of how one organism can protect against a common pathogen [Figure 7.1].

In addition to direct inhibition, a commensal organism can mediate colonization resistance through activation of the immune system. Recently, Schieber *et al.* demonstrated how a commensal *Escherichia coli* strain protects against muscle wasting associated with gut trauma and/or infection (Schieber et al., 2015). By sequencing the microbial communities of mice with differential colitis severity, the authors identified an outgrowth of *Escherichia* species in the more resistant mice. *E. coli* isolate O21:H+ was subsequently isolated and administered to the susceptible mice, which were then protected from colitis-induced wasting. Notably, an unrelated commensal *E. coli* strain did not provide a protective effect. This strain specificity highlights the importance of using linked primary, rather than banked isolates, since strains of the same species can display extensive functional variation. Similarly, when mice were infected with *Salmonella* Typhimurium or *Burkholderia thailandensis*, precolonization with *E. coli* O21:H+ reduced the degree of wasting (Schieber et al., 2015). With additional experiments, the authors showed that this protective effect was not due to inhibition of pathogen colonization, but rather that commensal *E. coli* O21:H+, acting through the innate immune system, down-regulates muscle atrophy and promotes muscle regeneration. In this example, the pathogenesis of an infectious disease is altered because of a single commensal activating the immune system rather than a commensal directly inhibiting the growth of a pathogen.

The idea of distinct *E. coli* strains conferring colonization resistance is not new. In 1917, the *E. coli* strain Nissle was isolated from a soldier who did not develop diarrhea during an outbreak of shigellosis (Nissle, 1961). Since then, research on this probiotic strain has identified several mechanisms by which it outcompetes pathogens, including an efficient iron acquisition system (Sassone-Corsi and Raffatellu, 2015). Despite this early example of colonization resistance, previous updates to Koch's postulates have not considered the overall community context in which a pathogen does or does not induce a disease. As described in (Buffie et al., 2015; Schieber et al., 2015), the role of specific members of the microbial community in disease pathogenesis could only be identified with antibiotic treatment and subsequent microbial community sequencing, technical advances that Koch could not have imagined as nucleic acids and antibiotics were not discovered until years after his death.

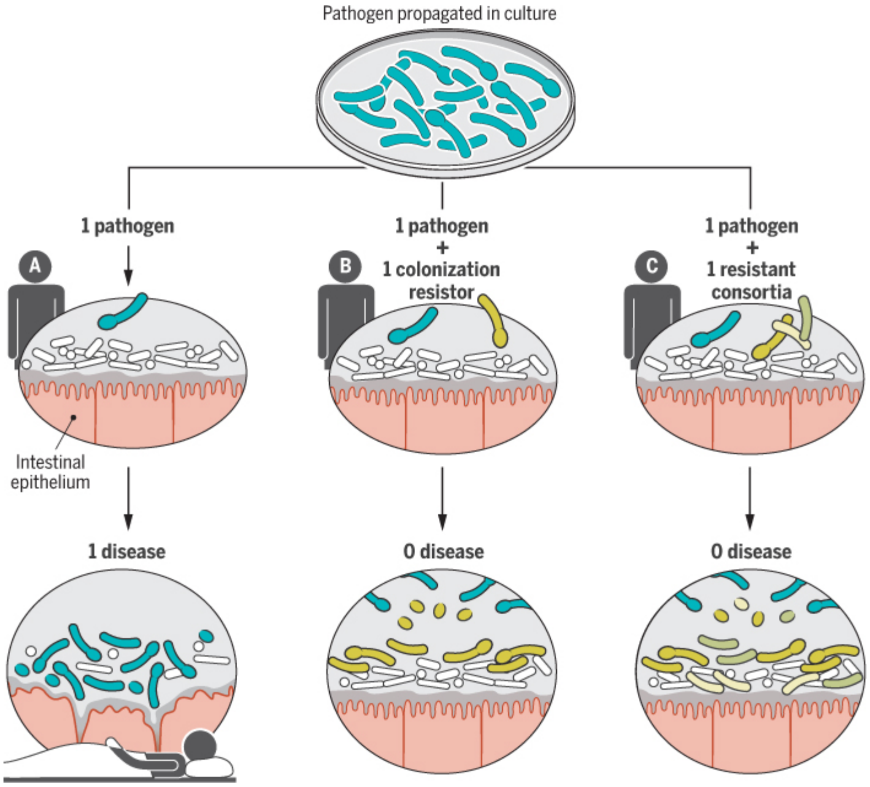


Figure 7.1. Microbial protectors

A) According to Koch's original postulates, a pathogenic organism in a host will induce disease. B) This assumption is challenged when an organism is present that can protect against the pathogen. C) In some cases, consortia of microbes can have an ever greater protective effect.

7.4 1 Pathogen + 1 Community = 0 Disease

The previous examples highlighted comparatively simple cases of disease causation in which single colonization resistors were important. However, multiple microbes can also have an enhanced protective effect.

As described above, Buffie *et al.* find that *C. scindens* provides colonization resistance to *C. difficile* (Buffie et al., 2015). However, the authors present evidence that even better outcomes are achieved when *C. scindens* cocolonizes with three other microbes. Similarly, Lawley *et al.* have reported that *C. difficile*-infected mice become less sick and clear the pathogen more efficiently after the administration of healthy donor feces (Lawley et al., 2012). To define a more tractable resistant community, Lawley *et al.* cultured individual isolates from the feces and combined them into phylogenetically distinct mixtures until they had found a six-member community that reproducibly reduced *C. difficile* infection and bacterial load.

These findings force us to consider under what circumstances a consortium of microbes can fulfill Koch's postulates. For example, do all members of the community have to be grown in pure culture and tested individually, or is it sufficient to grow and test a group culture? This is important in both scientific and translational arenas as researchers strive to create artificial communities capable of recapitulating the positive effects of fecal transplant in patients with recurrent *C. difficile* infections (van Nood et al., 2013).

In the future, artificial communities could also be created to treat other infections associated with antibiotic-induced alterations of the microbial community. For example, women taking antibiotics are prone to develop mucosal candidiasis due to a depletion of beneficial microbes (Break et al., 2015). Instead of traditional probiotic treatments, could a vaginal artificial community be designed to restore normal microbial community dynamics?

7.5 Toward dynamic adaption

Combining sequencing and culturing represents a powerful way to explore and define the microorganisms affecting disease outcome. With sequencing, all organisms present in a sample can be observed without the constraints imposed by pure culture requirements. Based on conclusions drawn and hypotheses generated from sequencing results, researchers can then proceed with a more targeted culturing approach to identify organisms of interest.

In summary, updated Koch's postulates incorporating sequencing and culturing would involve the following steps: first, sequencing to classify all members of the microbial community; second, using computational models to assess microbes both necessary and sufficient for disease induction; third, targeted culturing to isolate microbes of interest from the diseased host; and fourth, testing primary isolates and consortia in relevant disease models.

In regards to steps 3 and 4, it should be emphasized that working with primary isolates cultured directly from a diseased host, rather than commercially available strains, lends scientific accuracy given the extensive genetic variability within a species. As an additional step, testing nonprimary isolates can be done to evaluate how widespread the capability is across strains of a species.

In light of recent appreciation of microbial consortia, the scientific community should consider infectious disease causation in a broader systems biology context in which host genetic variability, health status, past exposure history, and microbial strains and communities are all important. As technology advances and new scientific discoveries are made, we must dynamically adapt Koch's postulates so today's science maintains the integrity that Koch originally fostered.

CHAPTER 8 BIBLIOGRAPHY

Alekseyenko, A.V., Perez-Perez, G.I., De Souza, A., Strober, B., Gao, Z., Bihan, M., Li, K., Methe, B.A., and Blaser, M.J. (2013). Community differentiation of the cutaneous microbiota in psoriasis. Microbiome *1*, 31.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Ames, S.K., Hysom, D.A., Gardner, S.N., Lloyd, G.S., Gokhale, M.B., and Allen, J.E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics *29*, 2253-2260.

Angiuoli, S.V., Matalka, M., Gussman, A., Galens, K., Vangala, M., Riley, D.R., Arze, C., White, J.R., White, O., and Fricke, W.F. (2011). CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC bioinformatics *12*, 356.

Asher, M.I., Keil, U., Anderson, H.R., Beasley, R., Crane, J., Martinez, F., Mitchell, E.A., Pearce, N., Sibbald, B., Stewart, A.W.*, et al.* (1995). International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. The European respiratory journal *8*, 483-491.

Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., Fukuda, S., Saito, T., Narushima, S., Hase, K.*, et al.* (2013). Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. Nature *500*, 232-236.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D.*, et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology *19*, 455-477.

Bantz, S.K., Zhu, Z., and Zheng, T. (2014). The Atopic March: Progression from Atopic Dermatitis to Allergic Rhinitis and Asthma. Journal of clinical & cellular immunology *5*.

Bardou, P., Mariette, J., Escudie, F., Djemiel, C., and Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. BMC bioinformatics *15*, 293.

Bath-Hextall, F.J., Birnie, A.J., Ravenscroft, J.C., and Williams, H.C. (2010). Interventions to reduce Staphylococcus aureus in the management of atopic eczema: an updated Cochrane review. The British journal of dermatology *163*, 12-26.

Belkaid, Y., and Hand, T.W. (2014). Role of the microbiota in immunity and inflammation. Cell *157*, 121-141.

Belkaid, Y., Piccirillo, C.A., Mendez, S., Shevach, E.M., and Sacks, D.L. (2002). CD4+CD25+ regulatory T cells control Leishmania major persistence and immunity. Nature *420*, 502-507.

Belkaid, Y., and Segre, J.A. (2014). Dialogue between skin microbiota and immunity. Science *346*, 954-959.

Belkaid, Y., and Tamoutounour, S. (2016). The influence of skin microorganisms on cutaneous immunity. Nature reviews Immunology *16*, 353-366.

Bhaduri, A., Qu, K., Lee, C.S., Ungewickell, A., and Khavari, P.A. (2012). Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics *28*, 1174-1175.

Bibby, K. (2013). Metagenomic identification of viral pathogens. Trends in biotechnology *31*, 275-279.

Blaser, M.J. (2014). The microbiome revolution. The Journal of clinical investigation *124*, 4162-4165.

Bode, L.G., Kluytmans, J.A., Wertheim, H.F., Bogaers, D., Vandenbroucke-Grauls, C.M., Roosendaal, R., Troelstra, A., Box, A.T., Voss, A., van der Tweel, I*., et al.* (2010). Preventing surgical-site infections in nasal carriers of Staphylococcus aureus. The New England journal of medicine *362*, 9-17.

Bogaert, D., Keijser, B., Huse, S., Rossen, J., Veenhoven, R., van Gils, E., Bruin, J., Montijn, R., Bonten, M., and Sanders, E. (2011). Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. PloS one *6*, e17035.

Bomar, L., Brugger, S.D., Yost, B.H., Davies, S.S., and Lemon, K.P. (2016). Corynebacterium accolens Releases Antipneumococcal Free Fatty Acids from Human Nostril and Skin Surface Triacylglycerols. mBio *7*, e01725-01715.

Brady, A., and Salzberg, S.L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature methods *6*, 673-676.

Break, T.J., Jaeger, M., Solis, N.V., Filler, S.G., Rodriguez, C.A., Lim, J.K., Lee, C.C., Sobel, J.D., Netea, M.G., and Lionakis, M.S. (2015). CX3CR1 is dispensable for control of mucosal Candida albicans infections in mice and humans. Infection and immunity *83*, 958-965.

Bruggemann, H., Henne, A., Hoster, F., Liesegang, H., Wiezer, A., Strittmatter, A., Hujer, S., Durre, P., and Gottschalk, G. (2004). The complete genome sequence of Propionibacterium acnes, a commensal of human skin. Science *305*, 671-673.

Buffie, C.G., Bucci, V., Stein, R.R., McKenney, P.T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., Viale, A*., et al.* (2015). Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature *517*, 205-208.

Buffie, C.G., and Pamer, E.G. (2013). Microbiota-mediated colonization resistance against intestinal pathogens. Nature reviews Immunology *13*, 790-801.

Butler-Wu, S.M., Sengupta, D.J., Kittichotirat, W., Matsen, F.A., 3rd, and Bumgarner, R.E. (2011). Genome sequence of a novel species, Propionibacterium humerusii. Journal of bacteriology *193*, 3678.

Byrd, A.L., Perez-Rogers, J.F., Manimaran, S., Castro-Nallar, E., Toma, I., McCaffrey, T., Siegel, M., Benson, G., Crandall, K.A., and Johnson, W.E. (2014). Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC bioinformatics *15*, 262.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I*., et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. Nature methods *7*, 335-336.

Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N*., et al.* (2011). Moving pictures of the human microbiome. Genome biology *12*, R50.

Chan, J.Z., Pallen, M.J., Oppenheim, B., and Constantinidou, C. (2012). Genome sequencing in clinical microbiology. Nature biotechnology *30*, 1068-1071.

Chan, J.Z., Sergeant, M.J., Lee, O.Y., Minnikin, D.E., Besra, G.S., Pap, I., Spigelman, M., Donoghue, H.D., and Pallen, M.J. (2013). Metagenomic analysis of tuberculosis in a mummy. The New England journal of medicine *369*, 289-290.

Chaptini, C., Quinn, S., and Marshman, G. (2016). Methicillin-resistant Staphylococcus aureus in children with atopic dermatitis from 1999 to 2014: A longitudinal study. The Australasian journal of dermatology 57(2),122-127.

Chen, E.C., Miller, S.A., DeRisi, J.L., and Chiu, C.Y. (2011a). Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. Journal of visualized experiments : JoVE (50), 2536.

Chen, E.C., Yagi, S., Kelly, K.R., Mendoza, S.P., Tarara, R.P., Canfield, D.R., Maninger, N., Rosenthal, A., Spinner, A., Bales, K.L*., et al.* (2011b). Cross-species transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. PLoS pathogens *7*, e1002155.

Chen, J.Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., and Tian, D. (2009). Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Molecular biology and evolution *26*, 1523-1531.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics *30*, 31-37.

Chin, C.S., Sorenson, J., Harris, J.B., Robins, W.P., Charles, R.C., Jean-Charles, R.R., Bullard, J., Webster, D.R., Kasarskis, A., Peluso, P*., et al.* (2011). The origin of the Haitian cholera outbreak strain. The New England journal of medicine *364*, 33-42.

Chng, K.R., Tay, A.S., Li, C., Ng, A.H., Wang, J., Suri, B.K., Matta, S.A., McGovern, N., Janela, B., Wong, X.F*., et al.* (2016). Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. Nature microbiology *1*, 16106.

Cho, S.H., Strickland, I., Boguniewicz, M., and Leung, D.Y. (2001). Fibronectin and fibrinogen contribute to the enhanced binding of Staphylococcus aureus to atopic skin. The Journal of allergy and clinical immunology *108*, 269-274.

Christensen, G.J., Scholz, C.F., Enghild, J., Rohde, H., Kilian, M., Thurmer, A., Brzuszkiewicz, E., Lomholt, H.B., and Bruggemann, H. (2016). Antagonism between Staphylococcus epidermidis and Propionibacterium acnes and its genomic basis. BMC genomics *17*, 152.

Chu, E.Y., Freeman, A.F., Jing, H., Cowen, E.W., Davis, J., Su, H.C., Holland, S.M., and Turner, M.L. (2012). Cutaneous manifestations of DOCK8 deficiency syndrome. Archives of dermatology *148*, 79-84.

Clark, R.A., Chong, B., Mirchandani, N., Brinster, N.K., Yamanaka, K., Dowgiert, R.K., and Kupper, T.S. (2006). The vast majority of CLA+ T cells are resident in normal skin. Journal of immunology *176*, 4431-4439.

CLSI (2013). Performance standards for antimicrobial susceptibility testing. CLSI approved standard M100-S23. (Wayne, PA.: Clinical and Laboratory Standards Institute).

Conlan, S., Mijares, L.A., Program, N.C.S., Becker, J., Blakesley, R.W., Bouffard, G.G., Brooks, S., Coleman, H., Gupta, J., Gurson, N*., et al.* (2012). Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome biology *13*, R64.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. Science *326*, 1694-1697.

Cullen, T.W., Schofield, W.B., Barry, N.A., Putnam, E.E., Rundell, E.A., Trent, M.S., Degnan, P.H., Booth, C.J., Yu, H., and Goodman, A.L. (2015). Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. Science *347*, 170-175.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic acids research *30*, 2478-2483.

DeLeo, F.R., Otto, M., Kreiswirth, B.N., and Chambers, H.F. (2010). Community-associated meticillin-resistant Staphylococcus aureus. Lancet *375*, 1557-1568.

Deng, Y.M., Caldwell, N., and Barr, I.G. (2011). Rapid detection and subtyping of human influenza A viruses and reassortants by pyrosequencing. PloS one *6*, e23400.

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E., and Crook, D.W. (2012). Transforming clinical microbiology with bacterial genome sequencing. Nature reviews Genetics *13*, 601-612.

Dimon, M.T., Wood, H.M., Rabbitts, P.H., and Arron, S.T. (2013). IMSA: integrated metagenomic sequence analysis for identification of exogenous reads in a host genomic background. PloS one *8*, e64546.

Donaldson, G.P., Lee, S.M., and Mazmanian, S.K. (2016). Gut biogeography of the bacterial microbiota. Nature reviews Microbiology *14*, 20-32.

Du Toit, G., Roberts, G., Sayre, P.H., Bahnson, H.T., Radulovic, S., Santos, A.F., Brough, H.A., Phippard, D., Basting, M., Feeney, M.*, et al.* (2015). Randomized trial of peanut consumption in infants at risk for peanut allergy. The New England journal of medicine *372*, 803-813.

Dunne, W.M., Jr., Westblade, L.F., and Ford, B. (2012). Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology *31*, 1719-1726.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635-1638.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460-2461.

Escherich, T. (1988). The intestinal bacteria of the neonate and breast-fed infant. 1884. Reviews of infectious diseases *10*, 1220-1225.

Escherich, T. (1989). The intestinal bacteria of the neonate and breast-fed infant. 1885. Reviews of infectious diseases *11*, 352-356.

Evans, A.S. (1976). Causation and disease: the Henle-Koch postulates revisited. The Yale journal of biology and medicine *49*, 175-195.

Eyerich, K., Eyerich, S., and Biedermann, T. (2015). The Multi-Modal Immune Pathogenesis of Atopic Eczema. Trends in immunology *36*, 788-801.

Faith, J.J., Colombel, J.F., and Gordon, J.I. (2015). Identifying strains that contribute to complex diseases through the study of microbial inheritance. Proceedings of the National Academy of Sciences of the United States of America *112*, 633-640.

Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L*., et al.* (2013). The long-term stability of the human gut microbiota. Science *341*, 1237439.

Falkow, S. (1988). Molecular Koch's postulates applied to microbial pathogenicity. Reviews of infectious diseases *10 Suppl 2*, S274-276.

Faveri, M., Mayer, M.P., Feres, M., de Figueiredo, L.C., Dewhirst, F.E., and Paster, B.J. (2008). Microbiological diversity of generalized aggressive periodontitis by 16S rRNA clonal analysis. Oral microbiology and immunology *23*, 112-118.

Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J.A., Schoenfeld, D., Nomicos, E., Park, M., Program, N.I.H.I.S.C.C.S*., et al.* (2013). Topographic diversity of fungal and bacterial communities in human skin. Nature *498*, 367-370.

Fitz-Gibbon, S., Tomida, S., Chiu, B.H., Nguyen, L., Du, C., Liu, M., Elashoff, D., Erfe, M.C., Loncaric, A., Kim, J*., et al.* (2013). Propionibacterium acnes strain populations in the human skin microbiome associated with acne. The Journal of investigative dermatology *133*, 2152-2160.

Flores, G.E., Caporaso, J.G., Henley, J.B., Rideout, J.R., Domogala, D., Chase, J., Leff, J.W., Vazquez-Baeza, Y., Gonzalez, A., Knight, R*., et al.* (2014). Temporal variability is a personalized feature of the human microbiome. Genome biology *15*, 531.

Francis, O.E., Bendall, M., Manimaran, S., Hong, C., Clement, N.L., Castro-Nallar, E., Snell, Q., Schaalje, G.B., Clement, M.J., Crandall, K.A*., et al.* (2013). Pathoscope: species identification and strain attribution with unassembled sequencing data. Genome research *23*, 1721-1729.

Frank, C., Werber, D., Cramer, J.P., Askar, M., Faber, M., an der Heiden, M., Bernard, H., Fruth, A., Prager, R., Spode, A*., et al.* (2011). Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. The New England journal of medicine *365*, 1771-1780.

Franzosa, E.A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X.C., and Huttenhower, C. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nature reviews Microbiology *13*, 360-372.

Fredricks, D.N., and Relman, D.A. (1996). Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. Clinical microbiology reviews *9*, 18-33.

Gaitanis, G., Magiatis, P., Hantschke, M., Bassukas, I.D., and Velegraki, A. (2012). The Malassezia genus in skin and systemic diseases. Clinical microbiology reviews *25*, 106-141.

Gajer, P., Brotman, R.M., Bai, G., Sakamoto, J., Schutte, U.M., Zhong, X., Koenig, S.S., Fu, L., Ma, Z.S., Zhou, X*., et al.* (2012). Temporal dynamics of the human vaginal microbiota. Science translational medicine *4*, 132ra152.

Galens, K., Orvis, J., Daugherty, S., Creasy, H.H., Angiuoli, S., White, O., Wortman, J., Mahurkar, A., and Giglio, M.G. (2011). The IGS Standard Operating Procedure for Automated Prokaryotic Annotation. Standards in genomic sciences *4*, 244-251.

Gallo, R.L., and Hooper, L.V. (2012). Epithelial antimicrobial defence of the skin and intestine. Nature reviews Immunology *12*, 503-516.

Gao, Z., Tseng, C.H., Pei, Z., and Blaser, M.J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. Proceedings of the National Academy of Sciences of the United States of America *104*, 2927-2932.

Genetics, E.A., Lifecourse Epidemiology Eczema, C., Australian Asthma Genetics, C., and Australian Asthma Genetics Consortium, A. (2015). Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nature genetics *47*, 1449-1456.

Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Greenblum, S., Carr, R., and Borenstein, E. (2015). Extensive strain-level copy-number variation across human gut microbiome species. Cell *160*, 583-594.

Greninger, A.L., Chen, E.C., Sittler, T., Scheinerman, A., Roubinian, N., Yu, G., Kim, E., Pillai, D.R., Guyard, C., Mazzulli, T*., et al.* (2010). A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS one *5*, e13381.

Gribbon, E.M., Cunliffe, W.J., and Holland, K.T. (1993). Interaction of Propionibacterium acnes with skin lipids in vitro. Journal of general microbiology *139*, 1745-1751.

Grice, E.A. (2015). The intersection of microbiome and host at the skin interface: genomic- and metagenomic-based insights. Genome research *25*, 1514-1520.

Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., Program, N.C.S., Bouffard, G.G., Blakesley, R.W., Murray, P.R., *et al.* (2009). Topographical and temporal diversity of the human skin microbiome. Science *324*, 1190-1192.

Grice, E.A., and Segre, J.A. (2011). The skin microbiome. Nature reviews Microbiology *9*, 244-253.

Group, N.H.W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., *et al.* (2009). The NIH Human Microbiome Project. Genome research *19*, 2317-2323.

Guindon, S., Delsuc, F., Dufayard, J.F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. Methods in molecular biology *537*, 113-137.

Hand, T.W., Dos Santos, L.M., Bouladoux, N., Molloy, M.J., Pagan, A.J., Pepper, M., Maynard, C.L., Elson, C.O., 3rd, and Belkaid, Y. (2012). Acute gastrointestinal infection induces long-lived microbiota-specific T cell responses. Science *337*, 1553-1556.

Hoeger, P.H., Lenz, W., Boutonnier, A., and Fournier, J.M. (1992). Staphylococcal skin colonization in children with atopic dermatitis: prevalence, persistence, and transmission of toxigenic and nontoxigenic strains. The Journal of infectious diseases *165*, 1064-1068.

Holland, K.T., Greenman, J., and Cunliffe, W.J. (1979). Growth of cutaneous propionibacteria on synthetic medium; growth yields and exoenzyme production. The Journal of applied bacteriology *47*, 383-394.

Holtgrewe, M. (2010). Mason – A Read Simulator for Second Generation Sequencing Data. Technical Report FU Berlin.

Hsiang, M.S., Shiau, R., Nadle, J., Chan, L., Lee, B., Chambers, H.F., and Pan, E. (2012). Epidemiologic Similarities in Pediatric Community-Associated Methicillin-Resistant and Methicillin-Sensitive Staphylococcus aureus in the San Francisco Bay Area. Journal of the Pediatric Infectious Diseases Society *1*, 200-211.

Huang, J.T., Abrams, M., Tlougan, B., Rademaker, A., and Paller, A.S. (2009). Treatment of Staphylococcus aureus colonization in atopic dermatitis decreases disease severity. Pediatrics *123*, e808-814.

Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207-214.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome research *17*, 377-386.

Hyman, R.W., Fukushima, M., Diamond, L., Kumm, J., Giudice, L.C., and Davis, R.W. (2005). Microbes on the human vaginal epithelium. Proceedings of the National Academy of Sciences of the United States of America *102*, 7952-7957.

Iebba, V., Totino, V., Gagliardi, A., Santangelo, F., Cacciotti, F., Trancassini, M., Mancini, C., Cicerone, C., Corazziari, E., Pantanella, F*., et al.* (2016). Eubiosis and dysbiosis: the two sides of the microbiota. The new microbiologica *39*, 1-12.

Ingham, E., Holland, K.T., Gowland, G., and Cunliffe, W.J. (1981). Partial purification and characterization of lipase (EC 3.1.1.3) from Propionibacterium acnes. Journal of general microbiology *124*, 393-401.

Iwase, T., Uehara, Y., Shinji, H., Tajima, A., Seo, H., Takada, K., Agata, T., and Mizunoe, Y. (2010). Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. Nature *465*, 346-349.

Janek, D., Zipperer, A., Kulik, A., Krismer, B., and Peschel, A. (2016). High Frequency and Diversity of Antimicrobial Activities Produced by Nasal Staphylococcus Strains against Bacterial Competitors. PLoS pathogens *12*, e1005812.

Joo, H.S., Fu, C.I., and Otto, M. (2016). Bacterial strategies of resistance to antimicrobial peptides. Philosophical transactions of the Royal Society of London Series B, Biological sciences *371*.

Kang, D., Shi, B., Erfe, M.C., Craft, N., and Li, H. (2015). Vitamin B12 modulates the transcriptome of the skin microbiota in acne pathogenesis. Science translational medicine *7*, 293ra103.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic acids research *33*, 511-518.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome research *12*, 656-664.

Kim, D.W., Park, J.Y., Park, K.D., Kim, T.H., Lee, W.J., Lee, S.J., and Kim, J. (2009). Are there predominant strains and toxins of Staphylococcus aureus in atopic dermatitis patients? Genotypic characterization and toxin determination of S. aureus isolated in adolescent and adult patients with atopic dermatitis. The Journal of dermatology *36*, 75-81.

Kobayashi, T., Glatz, M., Horiuchi, K., Kawasaki, H., Akiyama, H., Kaplan, D.H., Kong, H.H., Amagai, M., and Nagao, K. (2015). Dysbiosis and Staphylococcus aureus Colonization Drives Inflammation in Atopic Dermatitis. Immunity *42*, 756-766.

Koch, R. (1890). Uber bakteriologische Forschung Verhandlung des X In- ternationalen Medichinischen Congresses. Paper presented at: Xth International Congress of Medicine, Berlin (Hirschwald, Berlin).

Koga, C., Kabashima, K., Shiraishi, N., Kobayashi, M., and Tokura, Y. (2008). Possible pathogenic role of Th17 cells for atopic dermatitis. The Journal of investigative dermatology *128*, 2625-2630.

Kong, H.H., Oh, J., Deming, C., Conlan, S., Grice, E.A., Beatson, M.A., Nomicos, E., Polley, E.C., Komarow, H.D., Program, N.C.S.*, et al.* (2012). Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. Genome research *22*, 850-859.

Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., Matot, E., Jona, G., Harmelin, A., Cohen, N.*, et al.* (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science *349*, 1101-1106.

Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G., Getz, G., and Meyerson, M. (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature biotechnology *29*, 393-396.

Kuczynski, J., Lauber, C.L., Walters, W.A., Parfrey, L.W., Clemente, J.C., Gevers, D., and Knight, R. (2012). Experimental and analytical tools for studying the human microbiome. Nature reviews Genetics *13*, 47-58.

Kunz, B., Oranje, A.P., Labreze, L., Stalder, J.F., Ring, J., and Taieb, A. (1997). Clinical validation and guidelines for the SCORAD index: consensus report of the European Task Force on Atopic Dermatitis. Dermatology *195*, 10-19.

Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., Sekizuka, T., Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y.*, et al.* (2010). Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. PLoS one *5*, e10256.

Kuusela, P. (1978). Fibronectin binds to Staphylococcus aureus. Nature *276*, 718-720.

Laborel-Preneron, E., Bianchi, P., Boralevi, F., Lehours, P., Fraysse, F., Morice-Picard, F., Sugai, M., Sato'o, Y., Badiou, C., Lina, G.*, et al.* (2015). Effects of the Staphylococcus aureus and Staphylococcus epidermidis Secretomes Isolated from the Skin Microbiota of Atopic Children on CD4+ T Cell Activation. PLoS one *10*, e0141067.

Lai, Y., Di Nardo, A., Nakatsuji, T., Leichtle, A., Yang, Y., Cogen, A.L., Wu, Z.R., Hooper, L.V., Schmidt, R.R., von Aulock, S.*, et al.* (2009). Commensal bacteria regulate Toll-like receptor 3-dependent inflammation after skin injury. Nature medicine *15*, 1377-1382.

Lanciotti, R.S., Roehrig, J.T., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K.E., Crabtree, M.B., Scherret, J.H.*, et al.* (1999). Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. Science *286*, 2333-2337.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Langille, M.G., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R.*, et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology *31*, 814-821.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Lawley, T.D., Clare, S., Walker, A.W., Stares, M.D., Connor, T.R., Raisen, C., Goulding, D., Rad, R., Schreiber, F., Brandt, C.*, et al.* (2012). Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. PLoS pathogens *8*, e1002995.

Lax, S., Smith, D.P., Hampton-Marcell, J., Owens, S.M., Handley, K.M., Scott, N.M., Gibbons, S.M., Larsen, P., Shogan, B.D., Weiss, S.*, et al.* (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. Science *345*, 1048-1052.

Lee, S.M., Donaldson, G.P., Mikulski, Z., Boyajian, S., Ley, K., and Mazmanian, S.K. (2013). Bacterial colonization factors control specificity and stability of the gut microbiota. Nature *501*, 426-429.

Lee, Y.K., and Mazmanian, S.K. (2010). Has the microbiota played a critical role in the evolution of the adaptive immune system? Science *330*, 1768-1773.

Lefebure, T., and Stanhope, M.J. (2007). Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome biology *8*, R71.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: human gut microbes associated with obesity. Nature *444*, 1022-1023.

Leyden, J.J., Marples, R.R., and Kligman, A.M. (1974). Staphylococcus aureus in the lesions of atopic dermatitis. The British journal of dermatology *90*, 525-530.

Leyden, J.J., McGinley, K.J., Mills, O.H., and Kligman, A.M. (1975). Propionibacterium levels in patients with and without acne vulgaris. The Journal of investigative dermatology *65*, 382-384.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987-2993.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

Lieberman, T.D., Flett, K.B., Yelin, I., Martin, T.R., McAdam, A.J., Priebe, G.P., and Kishony, R. (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nature genetics *46*, 82-87.

Lienau, E.K., Strain, E., Wang, C., Zheng, J., Ottesen, A.R., Keys, C.E., Hammack, T.S., Musser, S.M., Brown, E.W., Allard, M.W*., et al.* (2011). Identification of a salmonellosis outbreak by means of molecular sequencing. The New England journal of medicine *364*, 981-982.

Ling, L.L., Schneider, T., Peoples, A.J., Spoering, A.L., Engels, I., Conlon, B.P., Mueller, A., Schaberle, T.F., Hughes, D.E., Epstein, S*., et al.* (2015). A new antibiotic kills pathogens without detectable resistance. Nature *517*, 455-459.

Lomholt, H., Andersen, K.E., and Kilian, M. (2005). Staphylococcus aureus clonal dynamics and virulence factors in children with atopic dermatitis. The Journal of investigative dermatology *125*, 977-982.

Luckey, T.D. (1972). Introduction to intestinal microecology. The American journal of clinical nutrition *25*, 1292-1294.

Macias, E.S., Pereira, F.A., Rietkerk, W., and Safai, B. (2011). Superantigens in dermatology. Journal of the American Academy of Dermatology *64*, 455-472; quiz 473-454.

Marples, R.R., Downing, D.T., and Kligman, A.M. (1971). Control of free fatty acids in human surface lipids by Corynebacterium acnes. The Journal of investigative dermatology *56*, 127-131.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 10-12.

Martineau, F., Picard, F.J., Roy, P.H., Ouellette, M., and Bergeron, M.G. (1996). Species-specific and ubiquitous DNA-based assays for rapid identification of Staphylococcus epidermidis. Journal of clinical microbiology *34*, 2888-2893.

McDowell, A., Barnard, E., Nagy, I., Gao, A., Tomida, S., Li, H., Eady, A., Cove, J., Nord, C.E., and Patrick, S. (2012). An expanded multilocus sequence typing scheme for propionibacterium acnes: investigation of 'pathogenic', 'commensal' and antibiotic resistant strains. PLoS One *7*, e41480.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.

Meisel, J.S., Hannigan, G.D., Tyldsley, A.S., SanMiguel, A.J., Hodkinson, B.P., Zheng, Q., and Grice, E.A. (2016). Skin Microbiome Surveys Are Strongly Influenced by Experimental Design. The Journal of investigative dermatology *136*, 947-956.

Mueller, N.T., Bakacs, E., Combellick, J., Grigoryan, Z., and Dominguez-Bello, M.G. (2015). The infant microbiome development: mom matters. Trends in molecular medicine *21*, 109-117.

Naeem, R., Rashid, M., and Pain, A. (2013). READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics *29*, 391-392.

Nagy, I., Pivarcsi, A., Kis, K., Koreck, A., Bodai, L., McDowell, A., Seltmann, H., Patrick, S., Zouboulis, C.C., and Kemeny, L. (2006). Propionibacterium acnes and lipopolysaccharide induce the expression of antimicrobial peptides and proinflammatory cytokines/chemokines in human sebocytes. Microbes and infection / Institut Pasteur *8*, 2195-2205.

Naik, S., Bouladoux, N., Linehan, J.L., Han, S.J., Harrison, O.J., Wilhelm, C., Conlan, S., Himmelfarb, S., Byrd, A.L., Deming, C.*, et al.* (2015). Commensal-dendritic-cell interaction specifies a unique protective skin immune signature. Nature *520*, 104-108.

Naik, S., Bouladoux, N., Wilhelm, C., Molloy, M.J., Salcedo, R., Kastenmuller, W., Deming, C., Quinones, M., Koo, L., Conlan, S.*, et al.* (2012). Compartmentalized control of skin immunity by resident commensals. Science *337*, 1115-1119.

Nakamura, Y., Oscherwitz, J., Cease, K.B., Chan, S.M., Munoz-Planillo, R., Hasegawa, M., Villaruz, A.E., Cheung, G.Y., McGavin, M.J., Travers, J.B.*, et al.* (2013).

Staphylococcus delta-toxin induces allergic skin disease by activating mast cells. Nature *503*, 397-401.

Nakatsuji, T., Chen, T.H., Two, A.M., Chun, K.A., Narala, S., Geha, R.S., Hata, T.R., and Gallo, R.L. (2016). Staphylococcus aureus exploits epidermal barrier defects in atopic dermatitis to trigger cytokine expression. The Journal of investigative dermatology.

Nakatsuji, T., Chiang, H.I., Jiang, S.B., Nagarajan, H., Zengler, K., and Gallo, R.L. (2013). The microbiome extends to subepidermal compartments of normal skin. Nature communications *4*, 1431.

Newell, E.W., and Davis, M.M. (2014). Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. Nature biotechnology *32*, 149-157.

Niebuhr, M., Gathmann, M., Scharonow, H., Mamerow, D., Mommert, S., Balaji, H., and Werfel, T. (2011). Staphylococcal alpha-toxin is a strong inducer of interleukin-17 in humans. Infection and immunity *79*, 1615-1622.

Nissle, A. (1961). [Old and new experiences on therapeutic successes by restoration of the colonic flora with mutaflor in gastrointestinal diseases]. Die Medizinische Welt *29-30*, 1519-1523.

Noda, S., Suarez-Farinas, M., Ungar, B., Kim, S.J., de Guzman Strong, C., Xu, H., Peng, X., Estrada, Y.D., Nakajima, S., Honda, T.*, et al.* (2015). The Asian atopic dermatitis phenotype combines features of atopic dermatitis and psoriasis with increased TH17 polarization. The Journal of allergy and clinical immunology *136*, 1254-1264.

Oh, J., Byrd, A.L., Deming, C., Conlan, S., Program, N.C.S., Kong, H.H., and Segre, J.A. (2014). Biogeography and individuality shape function in the human skin metagenome. Nature *514*, 59-64.

Oh, J., Byrd, A.L., Park, M., Program, N.C.S., Kong, H.H., and Segre, J.A. (2016). Temporal stability of the human skin microbiome. Cell *165*, 854-866.

Oh, J., Conlan, S., Polley, E.C., Segre, J.A., and Kong, H.H. (2012). Shifts in human skin and nares microbiota of healthy children and adults. Genome medicine *4*, 77.

Oh, J., Freeman, A.F., Program, N.C.S., Park, M., Sokolic, R., Candotti, F., Holland, S.M., Segre, J.A., and Kong, H.H. (2013). The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. Genome research *23*, 2103-2114.

Oranje, A.P., Glazenburg, E.J., Wolkerstorfer, A., and de Waard-van der Spek, F.B. (2007). Practical issues on interpretation of scoring atopic dermatitis: the SCORAD index, objective SCORAD and the three-item severity score. The British journal of dermatology *157*, 645-648.

Otto, M. (2009). Staphylococcus epidermidis--the 'accidental' pathogen. Nature reviews Microbiology *7*, 555-567.

Palmer, C.N., Irvine, A.D., Terron-Kwiatkowski, A., Zhao, Y., Liao, H., Lee, S.P., Goudie, D.R., Sandilands, A., Campbell, L.E., Smith, F.J.*, et al.* (2006). Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. Nature genetics *38*, 441-446.

Pamer, E.G. (2016). Resurrecting the intestinal microbiota to combat antibiotic-resistant pathogens. Science *352*, 535-538.

Pasparakis, M., Haase, I., and Nestle, F.O. (2014). Mechanisms regulating skin immunity and inflammation. Nature reviews Immunology *14*, 289-301.

Patil, K.R., Haider, P., Pope, P.B., Turnbaugh, P.J., Morrison, M., Scheffer, T., and McHardy, A.C. (2011). Taxonomic metagenome sequence assignment with structured output models. Nature methods *8*, 191-192.

Paulino, L.C., Tseng, C.H., Strober, B.E., and Blaser, M.J. (2006). Molecular analysis of fungal microbiota in samples from healthy human skin and psoriatic lesions. Journal of clinical microbiology *44*, 2933-2941.

Peleg, A.Y., Hogan, D.A., and Mylonakis, E. (2010). Medically important bacterial-fungal interactions. Nature reviews Microbiology *8*, 340-349.

Picardo, M., Ottaviani, M., Camera, E., and Mastrofrancesco, A. (2009). Sebaceous gland lipids. Dermato-endocrinology *1*, 68-71.

PrabhuDas, M., Adkins, B., Gans, H., King, C., Levy, O., Ramilo, O., and Siegrist, C.A. (2011). Challenges in infant immunity: implications for responses to infection and vaccines. Nature immunology *12*, 189-194.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D.*, et al.* (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature biotechnology *29*, 742-749.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic acids research *40*, D130-135.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rasko, D.A., Webster, D.R., Sahl, J.W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E.E., Sebra, R., Chin, C.S., Iliopoulos, D*., et al.* (2011). Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. The New England journal of medicine *365*, 709-717.

Relman, D.A., and Falkow, S. (2001). The meaning and impact of the human genome sequence for microbiology. Trends in microbiology *9*, 206-208.

Rivers, T.M. (1937). Viruses and Koch's Postulates. Journal of bacteriology *33*, 1-12.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., Chen, W., Pu, F., Peng, Y., Li, J*., et al.* (2011). Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. The New England journal of medicine *365*, 718-724.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M*., et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348-352.

Salipante, S.J., SenGupta, D.J., Cummings, L.A., Land, T.A., Hoogestraat, D.R., and Cookson, B.T. (2015). Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. Journal of clinical microbiology *53*, 1072-1079.

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., and Belshaw, R. (2010). Viral mutation rates. Journal of virology *84*, 9733-9748.

Sarkar, I.N., Planet, P.J., and Desalle, R. (2008). caos software for use in character-based DNA barcoding. Molecular ecology resources *8*, 1256-1259.

Sassone-Corsi, M., and Raffatellu, M. (2015). No vacancy: how beneficial microbes cooperate with immunity to provide colonization resistance to pathogens. Journal of immunology *194*, 4081-4087.

Savage, D.C. (1977). Microbial ecology of the gastrointestinal tract. Annual review of microbiology *31*, 107-133.

Scharschmidt, T.C., and Fischbach, M.A. (2013). What Lives On Our Skin: Ecology, Genomics and Therapeutic Opportunities Of the Skin Microbiome. Drug discovery today Disease mechanisms *10*.

Scharschmidt, T.C., List, K., Grice, E.A., Szabo, R., Program, N.C.S., Renaud, G., Lee, C.C., Wolfsberg, T.G., Bugge, T.H., and Segre, J.A. (2009). Matriptase-deficient mice exhibit ichthyotic skin with a selective shift in skin microbiota. The Journal of investigative dermatology *129*, 2435-2442.

Scharschmidt, T.C., Vasquez, K.S., Truong, H.A., Gearty, S.V., Pauli, M.L., Nosbaum, A., Gratz, I.K., Otto, M., Moon, J.J., Liese, J*., et al.* (2015). A Wave of Regulatory T Cells into Neonatal Skin Mediates Tolerance to Commensal Microbes. Immunity *43*, 1011-1021.

Schieber, A.M., Lee, Y.M., Chang, M.W., Leblanc, M., Collins, B., Downes, M., Evans, R.M., and Ayres, J.S. (2015). Disease tolerance mediated by microbiome E. coli involves inflammasome and IGF-1 signaling. Science *350*, 558-563.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J*., et al.* (2013). Genomic variation landscape of the human gut microbiome. Nature *493*, 45-50.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J*., et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology *75*, 7537-7541.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863-864.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Fungal Barcoding, C., and Fungal Barcoding Consortium Author, L. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences of the United States of America *109*, 6241-6246.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods *9*, 811-814.

Segre, J.A. (2006). Epidermal barrier formation and recovery in skin disorders. The Journal of clinical investigation *116*, 1150-1158.

Sender, R., Fuchs, S., and Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell *164*, 337-340.

Snitkin, E.S., Zelazny, A.M., Thomas, P.J., Stock, F., Group, N.C.S.P., Henderson, D.K., Palmore, T.N., and Segre, J.A. (2012). Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Science translational medicine *4*, 148ra116.

Sorg, J.A., and Sonenshein, A.L. (2010). Inhibiting the initiation of Clostridium difficile spore germination using analogs of chenodeoxycholic acid, a bile acid. Journal of bacteriology *192*, 4983-4990.

Stajich, J.E., Harris, T., Brunk, B.P., Brestelli, J., Fischer, S., Harb, O.S., Kissinger, J.C., Li, W., Nayak, V., Pinney, D.F.*, et al.* (2012). FungiDB: an integrated functional genomics database for fungi. Nucleic acids research *40*, D675-681.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688-2690.

Strange, P., Skov, L., Lisby, S., Nielsen, P.L., and Baadsgaard, O. (1996). Staphylococcal enterotoxin B applied on intact normal and intact atopic skin induces dermatitis. Archives of dermatology *132*, 27-33.

Suarez-Farinas, M., Dhingra, N., Gittler, J., Shemer, A., Cardinale, I., de Guzman Strong, C., Krueger, J.G., and Guttman-Yassky, E. (2013). Intrinsic atopic dermatitis shows similar TH2 and higher TH17 immune activation compared with extrinsic atopic dermatitis. The Journal of allergy and clinical immunology *132*, 361-370.

Suffia, I.J., Reckling, S.K., Piccirillo, C.A., Goldszmid, R.S., and Belkaid, Y. (2006). Infected site-restricted Foxp3+ natural regulatory T cells are specific for microbial antigens. The Journal of experimental medicine *203*, 777-788.

Suh, L., Coffin, S., Leckerman, K.H., Gelfand, J.M., Honig, P.J., and Yan, A.C. (2008). Methicillin-resistant Staphylococcus aureus colonization in children with atopic dermatitis. Pediatric dermatology *25*, 528-534.

Tomida, S., Nguyen, L., Chiu, B.H., Liu, J., Sodergren, E., Weinstock, G.M., and Li, H. (2013). Pan-genome and comparative genome analyses of propionibacterium acnes reveal its genomic diversity in the healthy and diseased human skin microbiome. mBio *4*, e00003-00013.

Tong, P.L., Roediger, B., Kolesnikoff, N., Biro, M., Tay, S.S., Jain, R., Shaw, L.E., Grimbaldeston, M.A., and Weninger, W. (2015a). The skin immune atlas: three-dimensional analysis of cutaneous leukocyte subsets by multiphoton microscopy. The Journal of investigative dermatology *135*, 84-93.

Tong, S.Y., Schaumburg, F., Ellington, M.J., Corander, J., Pichon, B., Leendertz, F., Bentley, S.D., Parkhill, J., Holt, D.C., Peters, G.*, et al.* (2015b). Novel staphylococcal species that form part of a Staphylococcus aureus-related complex: the non-pigmented Staphylococcus argenteus sp. nov. and the non-human primate-associated Staphylococcus schweitzeri sp. nov. International journal of systematic and evolutionary microbiology *65*, 15-22.

Torok, M.E., and Peacock, S.J. (2012). Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory--pipe dream or reality? The Journal of antimicrobial chemotherapy *67*, 2307-2308.

Totte, J.E., van der Feltz, W.T., Hennekam, M., van Belkum, A., van Zuuren, E.J., and Pasmans, S.G. (2016). Prevalence and odds of Staphylococcus aureus carriage in atopic dermatitis: a systematic review and meta-analysis. The British journal of dermatology.

Treangen, T.J., Koren, S., Sommer, D.D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A.E., Phillippy, A.M., and Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome biology *14*, R2.

Turner, M. (2011). Microbe outbreak panics Europe. Nature *474*, 137.

Ugolotti, E., Larghero, P., Vanni, I., Bandettini, R., Tripodi, G., Melioli, G., Di Marco, E., Raso, A., and Biassoni, R. (2016). Whole-genome sequencing as standard practice for the analysis of clonality in outbreaks of meticillin-resistant Staphylococcus aureus in a paediatric setting. The Journal of hospital infection *93*, 375-381.

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E.G., de Vos, W.M., Visser, C.E., Kuijper, E.J., Bartelsman, J.F., Tijssen, J.G.*, et al.* (2013). Duodenal infusion of donor feces for recurrent Clostridium difficile. The New England journal of medicine *368*, 407-415.

Volz, T., Skabytska, Y., Guenova, E., Chen, K.M., Frick, J.S., Kirschning, C.J., Kaesler, S., Rocken, M., and Biedermann, T. (2014). Nonpathogenic bacteria alleviating atopic dermatitis inflammation induce IL-10-producing dendritic cells and regulatory Tr1 cells. The Journal of investigative dermatology *134*, 96-104.

von Eiff, C., Becker, K., Machka, K., Stammer, H., and Peters, G. (2001). Nasal carriage as a source of Staphylococcus aureus bacteremia. Study Group. The New England journal of medicine *344*, 11-16.

Walker, M.J., and Beatson, S.A. (2012). Epidemiology. Outsmarting outbreaks. Science *338*, 1161-1162.

Webster, G.F., Ruggieri, M.R., and McGinley, K.J. (1981). Correlation of Propionibacterium acnes populations with the presence of triglycerides on nonhuman skin. Applied and environmental microbiology *41*, 1269-1270.

Weidenmaier, C., Goerke, C., and Wolz, C. (2012). Staphylococcus aureus determinants for nasal colonization. Trends in microbiology *20*, 243-250.

WHO (2004). The global burden of disease: 2004 update

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Use R, 1-212.

Williams, H.C., Burney, P.G., Pembroke, A.C., and Hay, R.J. (1994a). The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. III. Independent hospital validation. The British journal of dermatology *131*, 406-416.

Williams, H.C., Burney, P.G., Strachan, D., and Hay, R.J. (1994b). The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. II. Observer variation of clinical diagnosis and signs of atopic dermatitis. The British journal of dermatology *131*, 397-405.

Williams, M.R., and Gallo, R.L. (2015). The role of the skin microbiome in atopic dermatitis. Current allergy and asthma reports *15*, 65.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America *74*, 5088-5090.

Wollenberg, M.S., Claesen, J., Escapa, I.F., Aldridge, K.L., Fischbach, M.A., and Lemon, K.P. (2014). Propionibacterium-produced coproporphyrin III induces Staphylococcus aureus aggregation and biofilm formation. mBio *5*, e01286-01214.

Wu, G., Zhao, H., Li, C., Rajapakse, M.P., Wong, W.C., Xu, J., Saunders, C.W., Reeder, N.L., Reilman, R.A., Scheynius, A.*, et al.* (2015). Genus-Wide Comparative Genomics of Malassezia Delineates Its Phylogeny, Physiology, and Niche Adaptation on Human Skin. PLoS genetics *11*, e1005614.

Wylie, K.M., Mihindukulasuriya, K.A., Sodergren, E., Weinstock, G.M., and Storch, G.A. (2012). Sequence analysis of the human virome in febrile and afebrile children. PloS One *7*, e27735.

Xu, Y., Stange-Thomann, N., Weber, G., Bo, R., Dodge, S., David, R.G., Foley, K., Beheshti, J., Harris, N.L., Birren, B.*, et al.* (2003). Pathogen discovery from human tissue by sequence-based computational subtraction. Genomics *81*, 329-335.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., Sun, L., Zhang, T., Hu, Y., Du, J.*, et al.* (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. Journal of clinical microbiology *49*, 3463-3469.

Yassour, M., Vatanen, T., Siljander, H., Hamalainen, A.M., Harkonen, T., Ryhanen, S.J., Franzosa, E.A., Vlamakis, H., Huttenhower, C., Gevers, D.*, et al.* (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Science translational medicine *8*, 343ra381.

Yeung, M., Balma-Mena, A., Shear, N., Simor, A., Pope, E., Walsh, S., and McGavin, M.J. (2011). Identification of major clonal complexes and toxin producing strains among Staphylococcus aureus associated with atopic dermatitis. Microbes and infection / Institut Pasteur *13*, 189-197.

Yue, J.C., and Clayton, M.K. (2005). A similarity measure based on species proportions. Communications in Statistics-Theory and Methods *34*, 2123-2131.

Zhang, K., Sparling, J., Chow, B.L., Elsayed, S., Hussain, Z., Church, D.L., Gregson, D.B., Louie, T., and Conly, J.M. (2004). New quadriplex PCR assay for detection of methicillin and mupirocin resistance and simultaneous discrimination of Staphylococcus aureus from coagulase-negative staphylococci. Journal of clinical microbiology *42*, 4947-4955.

Zhang, L.J., Guerrero-Juarez, C.F., Hata, T., Bapat, S.P., Ramos, R., Plikus, M.V., and Gallo, R.L. (2015). Innate immunity. Dermal adipocytes protect against invasive Staphylococcus aureus skin infection. Science *347*, 67-71.

Zhu, A., Sunagawa, S., Mende, D.R., and Bork, P. (2015). Inter-individual differences in the gene content of human gut bacterial species. Genome biology *16*, 82.

Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics *29*, 2669-2677.

Zipperer, A., Konnerth, M.C., Laux, C., Berscheid, A., Janek, D., Weidenmaier, C., Burian, M., Schilling, N.A., Slavetinsky, C., Marschal, M*., et al.* (2016). Human commensals producing a novel antibiotic impair pathogen colonization. Nature *535*, 511-516.

CHAPTER 9 CURRICULUM VITAE

Allyson L. Byrd

[redacted]

Bioinformatics PhD Student |National Institutes of Health (NIH) | Boston University

EDUCATION

Boston University -National Institutes of Health Graduate Partnership Program
Expected October 2016
Boston, MA -Bethesda, MD
PhD in Bioinformatics
Development of pipelines for strain-level analysis of metagenomic microbiome data

University of Georgia, Honors College
May 2012
Athens, GA
Bachelor of Science in Genetics
G.P.A 3.95/4.00; *Summa cum laude with highest honors*

RESEARCH EXPERIENCE

PhD Dissertation *August 2013 - Present*
Bacterial strain tracking across the human skin landscape in health and disease *Bethesda, MD*
Advisors: Julie Segre, PhD – National Human Genome Research Institute (NHGRI)
Yasmine Belkaid, PhD – National Institute of Allergy and Infectious Diseases (NIAID)

- Designed and implemented novel pipelines to detect organisms in metagenomic samples of the skin microbiome
- Through collaborations with bioinformaticians, clinicians, and basic researchers, gained experience in effective communication, collaboration, and leadership of a multidisciplinary group
- Translated results from computational analyzes into hypotheses that were tested in murine models
- Presented results to audiences of varying expertise throughout the NIH and internationally
- Generated creative visualizations to display multifactorial microbiome and clinical data

Bioinformatics Challenge Project – Boston University *September 2012 – May 2013*
Rapid and accurate pathogen identification in unassembled sequencing data *Boston, MA*
Advisor: W. Evan Johnson, PhD – Division of Computational Biomedicine

- Worked collaboratively with a team of graduate students to develop the Clinical Pathoscope pipeline for pathogen detection in sequenced clinical samples

Rotation– Boston University *October 2012 – May 2013*
Comparative genomic analyzes of transcriptome assemblies *Boston, MA*
Advisor: John Finnerty, PhD – Department of Biology

- Designed a pipeline to detect the presence or absence of gene families in transcriptome assemblies
- Created the MySQL Calabase database for field researchers to log sightings of animal and plant species

Undergraduate Research – University of Georgia *October 2010 – May 2012*
Comparative genomics of Apicomplexa parasites *Athens, GA*
Advisor: Jessica Kissinger, PhD – The Center for Tropical and Emerging Global Diseases

- Identified and classified repeats in the phylum Apicomplexa as part of a group bioinformatics project
- Characterizing the ribosomal RNAs shared across Plasmodia for honors thesis

TECHNICAL STRENGTHS

Programming Languages	Python, R
Databases	MySQL
Bioinformatics	Next generation sequencing analysis, metagenomic sequence analysis, computational pipeline development, tuxedo suite, FlowJo
Communication	Microsoft Office, Adobe Illustrator

WORK EXPERIENCE

Oxford University Press *December 2012 – Present*
Technical Reviewer

- Consulted on over 100 bioinformatics web server and stand alone programs for possible publication in the Nucleic Acids Research annual Web Server issue
- Reviewed and made recommendation for webservers based on usability, novelty, and scientific impact

Marine Biological Laboratory *June – August 2012*
Course Assistant for Biology of Parasitism *Woods Hole, MA*

- Experienced complete scientific immersion, while performing administrative tasks for the lab and organizing outings for the students

FELLOWSHIPS & AWARDS

NIH Intramural Research Training Award (IRTA) *August 2012 – Present*

American Society of Microbiology Student Travel Grant *September 2016*

- Travel scholarship *Seattle, WA*

NIH Three Minute Talk (TmT) Contest *June 2016*

- 1st place *Bethesda, MD*

Boston University Bioinformatics Student Organized Symposium *June 2016*
- 1st place e-poster prize *Boston, MA*

NIH Graduate Student Research Symposium *January 2016*
- Poster award and travel scholarship *Bethesda, MD*

National Human Genome Research Institute Scientific Symposium *October 2015*
- Poster award *Bethesda, MD*

International Workshop on Bioinformatics and Systems Biology (ISMB) *July 2014*
- First place poster prize *Berlin, Germany*

International Workshop on Bioinformatics and Systems Biology (IBSB) *July 2013*
- F1000 Outstanding poster award and travel fellowship *Berlin, Germany*

PROFESSIONAL SERVICE AND LEADERSHIP

NIH Graduate Student Symposium Committee *January– December 2015*
Member *Bethesda, MD*
- Responsible for selecting keynote and student speakers for annual symposium

NIH Graduate Student Council *January– December 2014*
Co-chair *Bethesda, MD*
- Represented NIH graduate students on campus giving them a voice at the post-doc centric NIH
- Organized monthly meetings to oversee subcommittee progress and inform fellow students of upcoming events

NIH Graduate Student Retreat Committee *January– December 2014*
Member *Bethesda, MD*
- Organized the annual retreat by recruiting outside speakers and planning activities related to the theme "Science is a creative endeavor"
- Through active recruiting, superseded previous years registrations from 80 to 100 students

NIH Summer Intern Journal Club *June–July 2014*
Co-leader of *Exploring the World of Big Data with Computational Genomics* *Bethesda, MD*
- Introduced interns to reading and presenting scientific papers through an interactive series of lectures

Montgomery County Science Fair *March 2014,15,16*
Poster judge *Bethesda, MD*
- Provided feedback to middle and high school students on their science fair projects

PUBLICATIONS

- **Byrd, AL**, S Cassidy, OJ Harrison, C Deming, W Ng, NICS Program, JA Segre, and HH Kong. *Staphylococcal strain diversity underlies the individuality of atopic dermatitis*. Under Review.

- **Byrd, AL**†, J Oh†, M Park, NICS Program, HH Kong and JA Segre (2016). *Temporal Stability of the Human Skin Microbiome.* Cell 165, 854-866.

- **Byrd, AL** and JA Segre (2016). *Adapting Koch's postulates*. Science 351(6270):224-226.

- Morais da Fonseca, D†, TW Hand†, SJ Han, MY Gerner, A Glatman Zaretsky, **AL Byrd**, OJ Harrison, AM Ortiz, M Quinones, G Trinchieri, JM Brenchley, IE Brodsky, RN Germain, GJ Randolph, and Y Belkaid (2015). *Microbiota-dependent sequelae of acute infection compromises tissue specific immunity.* Cell 163, 354-366.

- Askenase, MH, SJ Han, **AL Byrd**, D Morais da Fonseca, N Bouladoux, C Wilhelm, JE Konkel, TW Hand, N Lacerda-Queiroz, XZ Su, G Trinchieri, JR Grainger and Y Belkaid (2015). *Bone-Marrow-Resident NK Cells Prime Monocytes for Regulatory Function during Infection.* Immunity 42(6): 1130-1142.

- **Byrd, AL** and JA Segre (2015). *Elucidating microbial codes to distinguish individuals.* Proc Natl Acad Sci U S A 112(22): 6778-6779.

- **Byrd, AL** and JA Segre (2015). *Integrating host gene expression and the microbiome to explore disease pathogenesis*. Genome Biol 16: 70.

- Naik, S†, N Bouladoux†, JL Linehan, SJ Han, OJ Harrison, C Wilhelm, S Conlan, S Himmelfarb, **AL Byrd**, C Deming, M Quinones, JM Brenchley, HH Kong, R Tussiwand, KM Murphy, M Merad, JA Segre and Y Belkaid (2015). *Commensal-dendritic-cell interaction specifies a unique protective skin immune signature.* Nature 520(7545): 104-108.

- Oh, J, **AL Byrd**, C Deming, S Conlan, NICS Program, HH Kong and JA Segre (2014). *Biogeography and individuality shape function in the human skin metagenome.* Nature 514(7520): 59-64.

- **Byrd, AL**†, JF Perez-Rogers†, S Manimaran, E Castro-Nallar, I Toma, T McCaffrey, M Siegel, G Benson, KA Crandall and WE Johnson (2014). *Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data.* BMC Bioinformatics 15: 262.

- Hong, C, S Manimaran, Y Shen, JF Perez-Rogers, **AL Byrd**, E Castro-Nallar, KA Crandall and WE Johnson (2014). *PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples.* Microbiome 2: 33.

- Stefanik, DJ†, TJ Lubinski†, BR Granger†, **AL Byrd**, AM Reitzel, L DeFilippo, A Lorenc and JR Finnerty (2014). *Production of a reference transcriptome and transcriptomic database (EdwardsiellaBase) for the lined sea anemone, Edwardsiella lineata, a parasitic cnidarian.* BMC Genomics 15: 71.

CONFERENCE, SEMINAR, & LECTURE PRESENTATIONS

Keystone Symposia: Immunity in Skin Development, Homeostasis, and Disease *Feb 2016*
Selected Speaker: *Tracking Strains of Staphylococci during Human Atopic Dermatitis Disease Progression* *Tahoe City, CA*

International Workshop on Bioinformatics and Systems Biology (IBSB) *July 2015*
Selected Speaker: *Bacterial strain tracking across the human skin landscape* *Boston, MA*

NIH FAES BIOL262: Research Tools for Studying Disease *April 2015*
Guest lecture: *Big Data Approaches: Analysis of next-generation sequencing data* *Bethesda, MD*

Boston University BF591: Applications in Translational Bioinformatics *March 2015*
Guest lecture: *Microbiome and metagenomics* *Boston, MA*

NIH Graduate Student Research Symposium *January 2015*
Selected Speaker: *Bacterial strain tracking across the human skin landscapeBethesda, MD*

International Workshop on Bioinformatics and Systems Biology (ISMB) *July 2014*
Selected Speaker: *Bacterial strain tracking across the human skin landscape* *Berlin, Germany*

Intelligent Systems for Molecular Biology (ISMB) *July 2013*
Selected Speaker: *Clinical PathoScope* *Berlin, Germany*

Center for Undergraduate Research Symposium, University of Georgia *May 2012*
Selected Speaker: *Comparative Genomics of Ribosomal RNAs in Malaria Parasites* *Athens, GA*

POSTER PRESENTATIONS

6th ASM Conference on Beneficial Microbes *September 2016*
Functional Implications of Staphylococcal Species in Atopic Dermatitis *Seattle, WA*

NIH Immunology Interest Group Retreat *September 2016*
Functional Implications of Staphylococcal Species in Atopic Dermatitis *Leesburg, VA*

Boston University Bioinformatics Student Organized Symposium *June 2016*
Staphylococcal strain diversity underlies the individuality of atopic dermatitis *Boston, MA*

Keystone Symposia: Immunity in Skin Development, Homeostasis, and Disease *Feb 2016*
Tracking Strains of Staphylococci during Human AD Disease Progression *Tahoe City, CA*

NHGRI Scientific Symposium *October 2015*
Temporal stability of the human skin microbiome *Bethesda, MD*

NIH Immunology Interest Group Retreat *September 2015*
Bacterial strain tracking across the human skin landscape demonstrates functional diversity within a species
Bethesda, MD

International Workshop on Bioinformatics and Systems Biology (IBSB) *July 2015*
Temporal stability of the human skin microbiome *Boston, MA*

Gut Microbiota Modulation of Host Physiology: The Search for Mechanism *March 2015*
Bacterial strain tracking across the human skin landscape *Keystone, CO*

NIH Graduate Student Research Symposium *January 2015*
Bacterial strain tracking across the human skin landscape *Bethesda, MD*

NHGRI Scientific Symposium *December 2014*
Bacterial strain tracking across the human skin landscape *Bethesda, MD*

NIH Immunology Interest Group Retreat *September 2014*
Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data *Bethesda, MD*

International Workshop on Bioinformatics and Systems Biology (IBSB) *July 2014*
Strain Tracking of Organisms in Metagenomic Samples of the Skin Microbiome *Berlin, Germany*

NIH Graduate Student Research Symposium *January 2014*
Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data *Bethesda, MD*

International Workshop on Bioinformatics and Systems Biology (IBSB) *July 2013*
Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data *Kyoto, Japan*

Intelligent Systems for Molecular Biology (ISMB) *July 2013*
Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data *Berlin, Germany*

Center for Undergraduate Research Symposium, University of Georgia *May 2012*
Comparative Genomics of Ribosomal RNAs in Malaria Parasites *Athens, GA*