Schmedes, Sarah E. Genetic Profiling of Skin Microbiomes for Forensic Human Identification. Doctor of Philosophy (Biomedical Sciences), September 2017, 162 pp., 12 tables, 27 figures, 130 references.

The field of microbial forensics has expanded from a focus in biodefense and biocrime attribution to include various metagenomics and microbiome applications made possible by advancements in sequencing and bioinformatics technologies. Recent developments in metagenomics and microbiome research with application to the forensic sciences, include post-mortem interval, body fluid identification, recent geolocation, and human identification. The primary goal of the dissertation described herein was to assess the feasibility of human identification from skin microbiomes using both shotgun metagenomic sequencing and targeted enrichment strategies. The main studies of this dissertation were conducted under the hypothesis that genes from stable, universal microbial species from the core skin microbiome can differentiate skin microbiomes of individuals and be applied towards forensic human identification purposes.

The initial study presented describes the development of a tool, AutoCurE, used to identify errors in bacterial genome metadata from public databases and curate the data for subsequent use in comparative genomic studies. This study highlights the types of inconsistencies and errors which may be present in public genome databases and describes the development of a curated local bacterial database for use in subsequent studies. This doctoral research herein presents the development of a novel approach for human identification using stable, universal clade-specific markers from skin microbiomes. Initially, publically available shotgun metagenomic datasets generated from skin microbiome samples collected from 17 body sites from 12 individuals, sampled over three time points over the course of ~3-year period, were mined to identify stable, universal microbial markers. Supervised learning, specifically regularized multinomial logistic regression and 1-nearest-neighbor classification, were performed using the nucleotide diversities of clade-specific markers to predict the correct classification of skin microbiomes to their respective host individuals. Reduced subsets of markers were developed into a novel targeted metagenomics sequencing panel, the hidSkinPlex, to generate individual-specific skin microbiome profiles to use for human identification. Finally, the hidSkinPlex was evaluated on skin microbiome samples collected from eight individuals and three body sites, in triplicate, to demonstrate a proof-of-concept to differentiate individuals with high accuracy.

The hidSkinPlex, comprised of 282 bacterial and 4 phage markers from 22 family-, genus-, species-, and subspecies-level clades, was used to correctly identify skin microbiomes from their respective donors with up to 92%, 96%, and 100% accuracy using samples from the foot, manubrium, and hand, respectively. Additionally, skin microbiomes were classified with up to 97% accuracy when the body site was unknown, and body site origin could be predicted with up to 86% accuracy. The hidSkinPlex is the first targeted metagenomics sequencing panel and method designed specifically for skin microbiomes with the intent of forensic human identification applications.

KEYWORDS: Bacteria · Genome database curation · Automation · AutoCurE · Skin microbiome · Human identification · Forensic profiling · Metagenomics · Supervised learning · Bioinformatics · Clade-specific marker · Targeted massively parallel sequencing · hidSkinPlex

GENETIC PROFILING OF SKIN MICROBIOMES FOR FORENSIC HUMAN IDENTIFICATION

DISSERTATION

Presented to the Graduate Council of the
Graduate School of Biomedical Sciences
University of North Texas
Health Science Center at Fort Worth
In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Sarah E. Schmedes, M.S.
Fort Worth, Texas
September, 2017

ACKNOWLEDGEMENTS

The past six years of my doctoral studies have been the most challenging and rewarding of my life. None of my success would have been possible without my colleagues, friends and family.

First and foremost, I would like to acknowledge my major professor Dr. Bruce Budowle. During my PhD studies, you have helped me grow into the scientist, writer, and professional I am today. I cannot thank you enough for the amazing opportunities you have granted me over the years – many of which have been once in a lifetime. You have always put your students first, pushing us to succeed, and have always encouraged us to challenge our own ideas and even yours. I consider myself beyond fortunate to have had you as my major professor, and I thank you for your selflessness and friendship. I have a true mentor for life.

I also would like to express my deep appreciation for my advisory committee Drs. Jianye Ge, Johnny He, Randall Murch, and Dana Kadavy, as well as my University Member, Dr. Robert Barber, for all your support and guidance over the past six years. You have challenged me and helped me grow, and I have enjoyed working with each of you as I approached this milestone.

I would like to thank the Graduate School of Biomedical Sciences support staff and, especially, Carla Lee Johnson for guiding me through the doctoral program. Thank you for answering millions of emails and helping with every aspect of my time at UNTHSC.

To all students, staff, and the numerous visiting scientists from near and far whom I've had the privilege of working with over the years, I wouldn't be where I am today without you all. I must thank Jonathan King in particular for his mad Visual Basic and Excel skills, for being my Perl buddy, and for his tireless work for the lab. I am honored to have worked by your side for the last eight years, and I will truly miss working beside you, my friend. For showing me that I have an inner computer nerd, I would like to thank Dr. August Woerner. Without your guidance and

instruction, I would not be moving forward in a career in bioinformatics. Thank you for making programming less intimidating and giving me the confidence to embrace new skills. It has been a true privilege to work with you, and I am honored to have gained another mentor and friend.

Most importantly, I would like to acknowledge my family and friends for standing by my side through the years. Thank you to my best friends for life Sarah Sheridan, Lora Ghobrial, Megan O'Donnell, and Rachel Keener for your words of advice and open ears. Thank you to my parents William and Katherine Bailey for always supporting me, teaching me to be a student of life, and giving me the confidence and courage to follow my dreams. Thank you to my awesome siblings Gregg and Jennifer, siblings-in-law Kelly, Josh, and Bradley, and my wonderful and amazing in-laws, Steve and Cari Chatman, for your encouragement. Thank you to my amazing husband Brian Chatman for being my number one fan, and helping me through the nights of panic – and for the ice cream runs when things got tough. These pages are not long enough to let you know how much you have supported me. Thank you for your sacrifice to help me realize my dreams and keep my training and career moving. And, finally, thank you to the cutest Cavalier King Charles Spaniel in the whole world, Bella, for providing endless cuddles during my studies, and for keeping my legs warm during countless hours of studying and writing.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

LIST OF TABLES

INTRODUCTION

Genetic Profiling of Skin Microbiomes for Forensic Human Identification

Microbial forensics traditionally has been defined as the use of scientific means to characterize microorganisms and their products to obtain attribution of a biological terrorist attack, biocrime, hoax, or accidental release of a biological agent (1). However, recent advancements in genome era technology, in particular massively parallel sequencing (MPS) and bioinformatics, have substantially expanded studies in microbial genomics, phylogenetics, and metagenomics. Because of enhanced genomics analysis capabilities, the focus of microbial forensics no longer concentrates solely on bioterrorism and biocrime but now can be extended to a more generalized definition of the use of scientific means to analyze microbial evidence to produce investigative leads in criminal and civil cases (2). The realm of microbial forensics now includes various applications using metagenomics and microbiome profiling for human identification, body fluid identification, post-mortem interval determination, material geolocation, and infection source tracking. Just a decade ago there were only about 300 sequenced prokaryotic genomes in publicly accessible databases (3). Today, more than 55,600 prokaryotic genomes have been sequenced at the finished, permanent draft, and draft status (4), and the number continues to increase rapidly. This increase of sequenced genomes in public databases allows for improved characterization of environmental metagenomes and microbiomes and detection of previously uncharacterized taxa (5).

The increased throughput and decreased cost of sequencing have enabled the completion of thousands of microbial genomes and metagenomes, resulting in an increased size and representation of microbial diversity in public genomic databases (e.g., National Center for Biotechnology Information (NCBI), Genbank (6); European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), European Nucleotide Archive (ENA) (7); and DNA Data Bank of Japan (DDBJ) (8). In fact, the goal of the Genomic Encyclopedia of Bacteria

and Archaea (GEBA) project was to expand the diversity of microbial species in databases (9–12), since databases were skewed by including only a small proportion of the known microbial diversity. Large-scale metagenomics studies, such as the Human Microbiome Project (HMP) (13, 14) and the Earth Microbiome Project (EMP) (15), were initiated to provide baseline data of microbial life in and on the human body and comprised within environmental ecosystems around the globe, respectively. Currently, more than 6,100 metagenome datasets are publically available on the Integrated Microbial Genomes & Microbiome Samples (IMG/M) databases (16). Numerous bioinformatics programs have been developed to support the analysis of single genomes and simple-to-complex metagenomes. This expansion of available genomic data, bioinformatics tools, and new technologies has given rise to new areas of research, in particular microbiome studies. The human microbiome has significant impacts on health, and more recently microbiome profiling has expanded into the forensic sciences. Microbiome profiling has been applied to a variety of forensic applications, including human identification (17–19), body fluid identification (20, 21), post-mortem interval (22, 23), recent geolocation (24), diet and health (25), and infection source tracking in the case of biocrimes (26–28) (Figure 1).

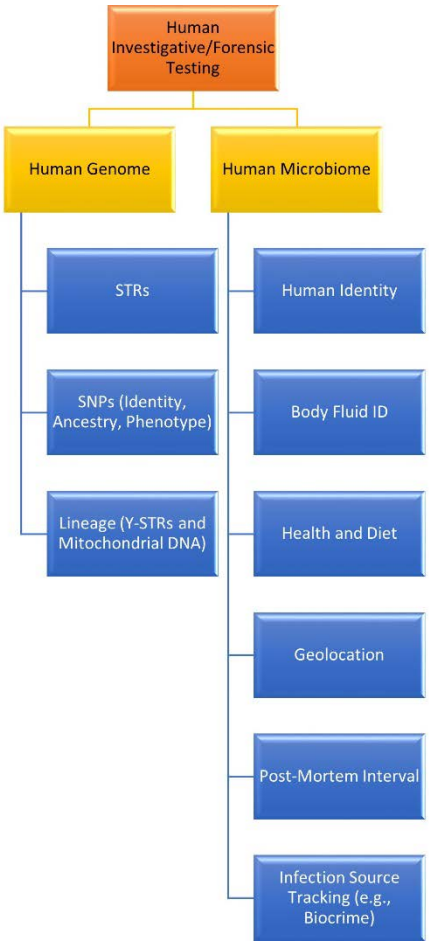

Figure 1. Expanded human and investigative forensic testing using human genome and human microbiome genetic markers (figure adapted from Schmedes et al. (2)).

The human microbiome is considered the second human genome (29) and harbors a vast diversity of microbial life in and on the surface of our bodies. The human microbiome is the collective group of microbial species, including bacteria, archaea, eukaryotes, and viruses, that inhabit the human body. Microbial cells outnumber human cells at a ratio of 10 to 1 (30), although that ratio has been suggested to be equal in cell number (31). Regardless, the number of microorganisms comprising the human microbiome is quite large. These vastly abundant microorganisms contribute more than 5,000,000 genes, from the gut alone, to complement the human repository of > 20,000 protein coding genes (14, 32). This increase in genetic complement

significantly contributes to essential bodily functions, such as metabolism, digestion, and immune response and plays a vital role in disease and health status (33). The NIH (National Institutes of Health) Human Microbiome Project (HMP) generated a baseline of healthy human microbiomes of various body sites, including the skin, nasal/respiratory tract, oral, gut, and urogenital tract, characterizing the taxonomic diversity and abundances of microbial species at each site (13, 14). The microbiome differs vastly in species composition and abundance at different areas of the body, and distinct microbial community signatures are specific to particular body sites (13, 14, 34). Human microbiomes consist of both relatively stable and transient microorganisms with changing abundances depending on various factors such as age, geography, diet, hygiene, health and antibiotic use (25, 35–38). Variation and alterations of the human microbiome also have been associated and linked with conditions such as obesity (25, 39), cancer (40), irritable bowel syndrome (41), metabolic syndrome (42), and bacterial vaginosis (43), to name a few. Notably, microbiomes have been shown to harbor microbial community signatures that differ among individuals (44), indicating that microbiomes could be highly individualizing and potentially unique to each individual. Thus, analysis of the human microbiome may be applicable to forensic purposes to gain intelligence information regarding a person's identity, recent geolocation, habitation, diet, and source of bodily fluid trace evidence, such as for touch DNA purposes.

Methods used to characterize the microbiome commonly adopt one of two approaches: targeted 16S rRNA and whole-genome shotgun sequencing (WGS). Depending on the method used, different types of data may be generated and inferred for microbiome characterization. Different characterization strategies employ the use of taxonomic classification of the whole microbial community, abundance ratios, testing alpha and beta diversity of the communities, functional gene content, and identification of specific genetic markers including antibiotic

resistance and virulence markers (14, 45, 46). The type of question to be answered will dictate the data obtained from a sample based on the metagenomic sequencing method used. The choice for using one method over another (i.e., targeted 16S rRNA and WGS) depends on the types of data required, throughput requirements and/or limitations, and cost.

The 16S rRNA gene encodes ribosomal RNA found in the small prokaryotic ribosomal subunit (30S). This locus is the most commonly used bacterial genetic marker in bacterial phylogenetic studies and broad bacterial identification. The conserved and variable regions of the gene, the available databases (e.g., the Ribosomal Database Project (47); Greengenes (48); SILVA (49)), and substantial volume of 16S rRNA studies add to the appeal of using this marker in a variety of applications. Targeted 16S rRNA metagenomic sequencing has been widely used in microbiome studies to study the taxonomic composition, taxa abundance ratios, and phylogenetic diversity within and among microbiomes (13, 14); however, the many limitations of using a single genetic marker cannot be ignored. The limitations to using solely 16S rRNA include insufficient genus or species resolution (50), PCR bias (51, 52), copy number variation (53) and sequence variability among a single bacterium (54), inaccurate phylogenetic relationships based on key variability outside of the marker region (55), and horizontal transfer of the entire gene region (56, 57). These phenomena can lead to inaccurate abundance ratios and taxonomic assignments.

WGS metagenomics sequencing is an alternative approach to that of targeting the 16S rRNA gene. The shotgun approach provides the theoretical ability to sequence the entire genome (DNA or RNA) of a single microorganism or an entire metagenome of many microorganisms in a given sample (Figure 2). Being more comprehensive in coverage, WGS metagenomics sequencing could provide species or strain level characterization, functional gene content, potential assembly of whole genomes, and identification of informative markers for antibiotic resistance and virulence

genes, which is not readily feasible by single marker analyses. However, WGS also has some limitations. The more area of any given genome that is covered, the less read depth will be obtained for any particular site, potentially reducing the confidence of a base call from sequence data and potentially missing informative sites for speciation, strain resolution or functionality studies. Therefore, the possibility of detecting species or strain-specific markers is reduced greatly. Highly complex metagenomic samples can contain thousands of species within a sample therefore making it difficult to obtain complete coverage of any one genome, especially those at low abundance. Depending on the complexity of the sample sequence reads generated for lower abundant species (and even high abundant species) may not be obtained or may be limited.

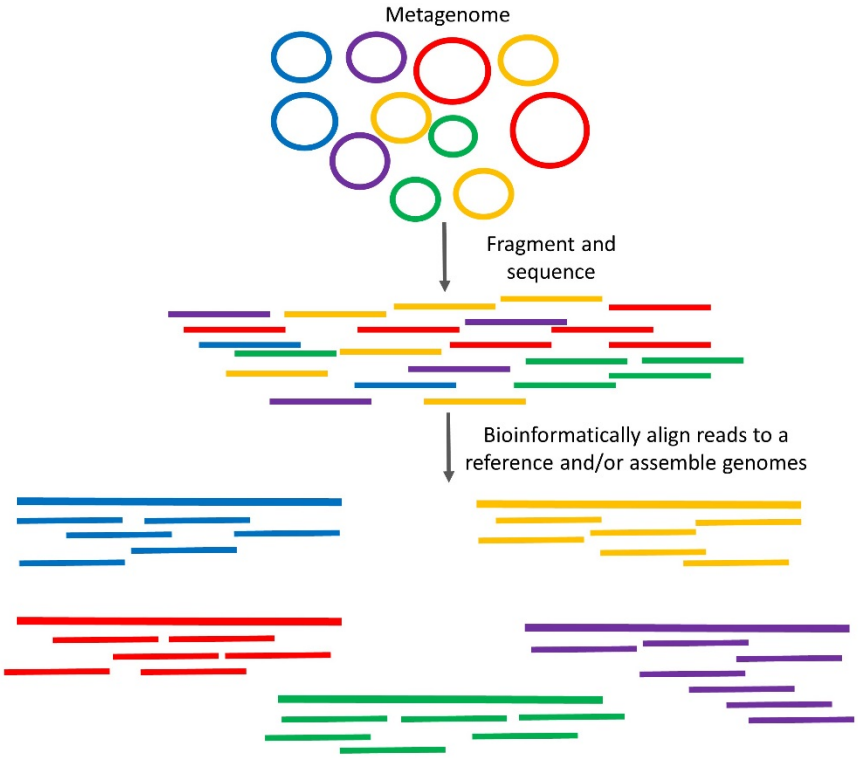

Figure 2. Schematic of WGS metagenomic sequencing. All genomic sequences within the metagenome are fragmented to universal lengths and sequenced in parallel. Bioinformatic methods are then used to align reads to reference genomes or assemble reads into contigs to reconstruct the community members of the metagenome.

Both targeted 16S rRNA and WGS metagenomics sequencing were used in the NIH HMP to generate > 5,000 metagenome datasets from 5 body regions (up to 18 body sites) from 242 healthy individuals (13, 14). The NIH HMP provided the sequence data and analysis tools to use these approaches for microbiome profiling for numerous applications beyond the scope of health and disease, such as the forensic sciences. Microbial forensic applications using metagenomics, including microbiome profiling, have focused on post-mortem interval (i.e., using microbial signatures from human decomposition to predict time-of-death) (22, 23), infection source tracking (e.g., determining patient zero in cases of deliberate or negligent transmission of HCV and HIV) (26–28), forensic identification of trace soil evidence (58), body fluid identification (20, 21), and human identification (17–19). The focus of this dissertation focuses on developing novel methods for forensic human identification using skin microbiomes.

Humans continually shed epithelial and microbial cells from skin surfaces onto touched items which leave traces of genetic material. The DNA from these cells, therefore, can be transferred via primary, secondary, and tertiary transfer onto other objects and surfaces. This concept is exploited for forensic purposes using human DNA typing to determine the identity of an individual(s) who may have touched an object at a crime scene. Current human forensic typing methods use a defined set of short tandem repeat (STR) (59) or single-nucleotide polymorphism (SNP) (60) markers to determine the identity of an individual based on the genetic profiles retrieved. However, in many cases the amount of DNA left behind on an object is too low (i.e.,

low copy number (LCN) DNA) to generate a complete genetic profile. Various methods have been used in LCN DNA typing to attempt to enhance the signal of a genetic profile using methods, which include sample dilution to reduce inhibition, sample concentration, increased number of polymerase chain reaction (PCR) cycles, whole-genome amplification (WGA), post-PCR purification, and increased injection times during capillary electrophoresis (CE) (61). However, each method has limitations and is susceptible to exacerbated stochastic effects. LCN typing methods and interpretation have been controversial at times and have had only limited success. Alternative methods using high-copy number markers (HCN), such as targeted hypervariable regions of the mitochondrial genome (62–64) (and soon to be whole mitochondrial genome (65)), are typically used in cases with highly degraded or LCN DNA, such as unidentified skeletal remains cases. In order to improve success of LCN typing one could employ the use of an orthogonal approach using another type of HCN marker(s), such as the use of microbial genetic typing in conjunction with human DNA typing. Since microbial cells outnumber human cells, (i.e., the typical number of bacterial cells from a single swab and scraping from a finger can range from ~10,000 bacteria/cm^2 to ~50,000 bacteria/cm^2, respectively (66)), it is plausible that microbial genetic profiling can be used alone or in conjunction with human DNA typing for forensic human identity purposes and potentially have a higher typing success rate and be a more robust assay. In addition, more information could be retrieved from microbial DNA profiles, such as potential recent geolocation and drug network associations (67).

The human skin microbiome (and virome) has been characterized, defining taxonomic composition and abundances, functional gene content, phylogenetic diversity, and temporal stability among various skin sites and among individuals (36, 66, 68–76). Microbial communities can vary vastly depending on the body environment sampled, such as dry, moist, or sebaceous

sites (68, 74, 76) (Figure 4). The skin microbiome is comprised of four dominant phyla: *Actinobacteria, Firmicutes*, *Proteobacteria*, and *Bacteroidetes*; and numerous other phyla have been detected in lower abundances (36, 68). As many as 19 phyla and 205 genera have been reported colonizing the skin (68), although these values vary depending on the study, sample cohort, and methodology used. The dominant genera of the skin microbiome include *Proprionibacterium*, *Staphylococcus*, and *Corynebacterium* (36, 68, 72, 74, 76, 77). Specific microflora of the skin have been identified that are associated with certain skin diseases, such as *Staphylococcus aureus*, a common pathogen associated with atopic dermatitis in children (78) and *Staphylococcus epidermidis*, a common commensal organism associated with nosocomial infections (79). Overall, the skin microbiome is more variable than the oral or gut microbiome (80). Although exposed externally, portions of the skin microbiome are highly stable and unique to an individual (76). Even after hand washing, microbial communities restore back to normal levels relatively quickly (36).

The idea that microbiomes are personal and unique to an individual has been supported to varying degrees. Meadow et al. (81) identified unique microbial clouds in the surrounding air within close proximity of specific persons. Franzosa et al. (19) identified stable personal metagenomic codes within individual microbiome samples from various body sites, using clade-specific markers and tiled kilobase windows compiled from bacterial reference genomes, to capture the strain-level variation to differentiate individuals. Due to the stability and unique signatures of the individual microbiome, one could exploit these signatures for human forensic identification. In fact, recent studies have demonstrated the potential to use skin microbiome profiles for forensic applications, by using unsupervised methods to demonstrate that touched

items resemble their donors (17, 82, 83). Few studies have utilized supervised approaches for the purposes of classification of skin microbiomes (19, 24, 84).

In the dissertation herein, both unsupervised and supervised learning (i.e., machine learning) were used to characterize skin microbiomes for the purposes of human identification. Within the context of the human microbiome, various applications of machine learning methods have been previously described (85). Briefly, unsupervised methods are primarily used for data visualization, and are conducted without utilizing information on the dependent variable (Figure 3). Supervised methods, on the other hand, are used for prediction, and utilize information on both the dependent and independent variables. The unsupervised methods used in this dissertation include principal components analysis (PCA) and maximum-likelihood phylogenies. PCA can be used to perform dimension reduction, so that high-dimensional data can be visualized in a lower dimensional space. PCA does this by finding orthogonal linear combinations of the independent variables that maximize the variance. These linear combinations can then be visualized to assess if distinct patterns emerge. Maximum-likelihood phylogenetic trees were used to visualize and estimate the evolutionary distance between gene sequences amongst samples. Neither of these approaches use information on who, or from where, our metagenomic samples derive.

Unlike unsupervised learning, supervised learning attempts to utilize information on both the dependent and independent variables. In the context of supervised learning, all variables, dependent and independent, are called features; features are combined into a feature vector, which represents a single observation. Supervised learning includes regression (including linear regression), which is used for predicting continuous variables, and classification, which is used to predict categorical variables. The two supervised methods applied in this dissertation include regularized multinomial logistic regression (RMLR) and 1-nearest-neighbor classification (1NN)

using the Euclidean distance (i.e., the shortest distance using a straight line between any two points). 1NN predicts the state of a categorical variable (e.g., an individual), to which the classifier is blind, by assigning it the label its closest point (i.e., its 1-nearest-neighbor) under the Euclidean distance function. RMLR predicts a categorical variable (e.g., an individual) using a multinomial logistic regression that has been regularized. The regularization used in this dissertation is ridge regularization, which minimizes the sum of squares of the coefficients for all features. The ridge value itself then sets the relative importance of minimizing the model error (i.e., deviance) relative to the magnitude of the sum of squares of the coefficients themselves. Regularization penalizes features with large magnitude coefficients to reduce error and prevent overfitting (i.e., training the model to work well on the training data and less well on new data, in which case the model is fit to both noise and signal in the data). Feature selection (attribute selection) also was used in this dissertation to identify subsets of features which provide similar prediction accuracies compared to using all features. Feature selection helps reduce noise and eliminate features that do not contribute to the performance of the classifier.

Cross validation was used to assess the accuracy of all classification methods. Cross-validation uses two data sets: a training set on which the model is built, and test set used to assess the accuracy of the model. In this dissertation leave-one-out cross validation (LOOCV) was used. With LOOCV n models are created each on n-1 different samples (n=sample size), and then accuracy is assessed on the single left out sample. Therefore, each test variable is removed from the training set prior to classification. This approach maximizes the size of the training set while still reducing the effects of overfitting.

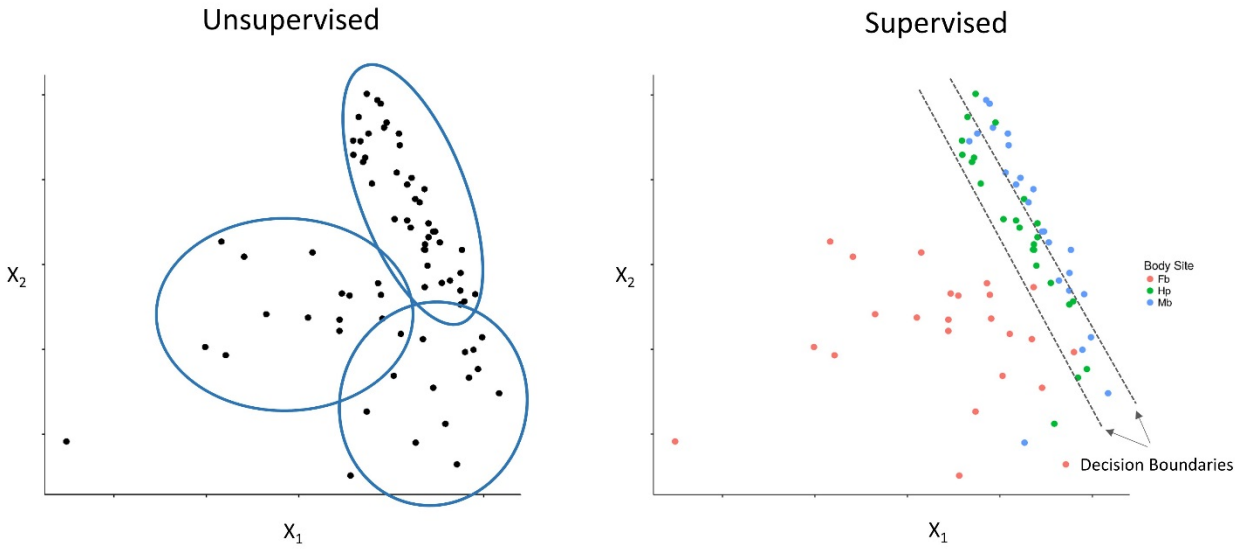

Figure 3. Comparison of supervised and unsupervised learning. The unsupervised method (left) is an example of a principal component analysis to visualize the maximized variance of the independent variables. Blue circles represent potential clusters observed indicating data may be correlated, which may or may not correspond to class labels. The supervised method (right) depicts the inferred hyperplanes computed by the classifier to separate data points based on the known class labels (e.g., body site). If an unknown test sample was introduced in the model, the classification assigned to the unknown variable would be in relation to the decision boundaries.

Recent studies using unsupervised methods have demonstrated that skin microbiome signatures detected from touched objects resemble signatures collected from their particular donors (17, 82, 83). Fierer et al. (17) demonstrated that skin-associated bacterial communities collected from touched objects, such as computer mice and keyboards, could be linked back to the owners. Goga (82) demonstrated that in most cases bacterial communities collected from shoes resembled skin bacterial communities of the wearers of the shoes. Few studies have utilized supervised approaches for prediction purposes of skin microbiomes (19, 24, 84). Franzosa et al. (19) used an implicit hitting set approach to identify minimum cardinality sets of presence/absence features, such as clade-specific markers and 1kb genomic windows, to identify strain-level metagenomics codes specific to individuals. More than 80% of individuals could be identified using codes from

gut microbiome samples; however, only ~30% of individuals could be identified using skin microbiomes (i.e., from anterior nares) sampled over 30-300 days. Lax et al. (24) and Williams et al. (84) applied random decision forests using operational taxonomic units (OTUs) from 16S rRNA sequences to differentiate individuals using skin microbiome samples. Lax et al. (24) performed a study of trace microbiome sampling from phones and shoes (and associated floor samples) as well as sampling from phones and shoes of individuals in three different geographical regions. Lax et al. (24) were able to associate skin microbiomes samples collected from phone surfaces (i.e., face and hand skin microbiome touch samples) to the owner of the phone with 96.3% accuracy; however, the majority of samples collected from each participant was only sampled at a single time point (24), a limitation also with the results in Williams et al. (84). Future studies utilizing supervised approaches for human identification, should demonstrate high classification accuracies using strain-level features to associate skin microbiomes to their individuals donors over long time intervals, making results more applicable to a typical forensic setting.

Various skin microbiome features have been used with unsupervised and supervised methods previously described above. Strain-level features from WGS metagenomic sequencing provide higher resolution than 16S rRNA based features, such as terminal restriction fragment length polymorphism profiles (82, 86, 87), OTUs abundances (18, 19, 24, 81, 83, 84), and biological community distances (e.g., UniFrac distance) (17, 24). The greatest temporal stability of strain-level features include *Propionibacterium acnes* single-nucleotide variant (SNV) profiles (76) and gene-level features, including clade-specific markers and 1kb genomic windows (19). Strain-level heterogeneity, measured by nucleotide diversity (as described by Nayfach et al. (88)), also has shown to be greater between individuals than within an individual (88). Strain-level features likely are most appropriate for human identification using skin microbiomes; however, a

method has yet to be described using supervised learning approaches with strain-level features stable over reasonably long time intervals.

Most studies characterizing skin microbiomes for forensic purposes have utilized targeted 16S rRNA or shotgun metagenomic sequencing; however, these methods are not ideal for forensic characterization of skin microbiomes due to limited species- and strain-level resolution and susceptibility of stochastic effects, respectively. An alternative metagenomics approach could use targeted enrichment of informative markers shown to provide individualizing resolution of skin microbiomes stable over time. A reliable method with the capability of strain-level resolution could be developed for forensic applications and allow for sufficient coverage of informative sites, even from body sites with low-abundant taxa. By developing a targeted microbiome profiling method, a more sensitive and specific typing method can be achieved for characterization of the microbiomes to associate to their human hosts. This method could provide an independent or an orthogonal approach that can be used in addition to standard human forensic typing methods. Additionally, this same approach can be used for other microbiome studies (i.e., gut, urogenital, respiratory, and oral) to develop typing methods for each microbiome body site relevant to forensics.

The overall goal of this doctoral research was to develop a novel metagenomics approach that targets select microbial species to characterize human skin microbiomes for forensic human identification. The doctoral dissertation presented herein was conducted under the hypothesis that genes from stable, universal microbial species from the core skin microbiome can differentiate skin microbiomes of individuals and be applied towards forensic human identification purposes. The specific aims of this doctoral research were: 1) to identify a set of universal, stable microbial genetic markers from skin microbiomes with the ability to differentiate individuals; 2) develop the

skin microbiome marker panel into a multiplex amplification assay to be used for targeted sequencing and an associated bioinformatics analysis pipeline; and 3) evaluate the marker panel and analysis pipeline on a subset of individuals to demonstrate proof-of-concept that targeted amplification and sequencing of a select subset of skin microbiome markers can be used for human identification purposes.

In the dissertation herein, results and findings from three studies are described. **Chapter 1**, "Correcting inconsistencies and errors in bacterial genome metadata using an automated curation tool in Excel (AutoCurE)" (Schmedes SE, King JL, and Budowle B. 2015. *Front Bioeng Biotechnol* 3:138), describes the development of an automated curation tool in Excel (AutoCurE) used to facilitate local genome database curation of bacterial genomes downloaded from NCBI. As more complete genomes are sequenced and become publically-available, there is a need for storing and maintaining quality-controlled local genome databases for comparative genomics studies. In this study, more than 2,700 publically-available bacterial genomes were downloaded to create a local bacterial genome database and inconsistencies and errors were identified in bacterial genome metadata associated with downloaded genomes. AutoCurE was developed to flag nine data fields related to genome report accession number, BioProject/UID consistency, accession number consistency, genus match, species match, identification of archaea, RefSeq reference genome accession, presence of chromosome/genome data, and identification of draft or partial genome sequence. Additional features of AutoCurE were designed to assist users with database and file manipulation to curate their local databases. AutoCurE provides an easy-to-use tool for Windows-based platforms for users without programming or advanced bioinformatics capabilities to generate a local bacterial genome database (sequence files) with quality control of associated

metadata for use for bacterial comparative genomic studies (See Appendix A and B for AutoCurE manuals).

Chapter 2, "Forensic human identification using skin microbiomes" (Schmedes SE, Woerner AE, and Budowle B. 2017. *Appl Environ Microbiol* (in press)), describes a novel approach to characterize skin microbiomes and use supervised learning to attribute skin microbial signatures to their respective individual hosts for potential forensic identification applications. Publically available shotgun metagenomic datasets generated from skin microbiome samples collected from 14 body sites from 12 healthy individuals for three time points over a ~3 year period were mined to identify stable microbial genetic markers which provide differentiating resolution of individuals. RMLR and 1NN were performed using two feature types derived from skin microbiomes signatures, *Propionibacterium acnes* pangenome gene presence/absence features and nucleotide diversity of universal clade-specific markers, to classify skin microbiomes to their individual donors. Classification accuracies computed using both feature types were assessed in a formal model testing framework to determine which feature type performed best for skin microbiome classification. Feature selection (i.e., identification of subsets of features which provides similar and/or greater power than using a full set of features) also was performed to identify a subset of features to include in a preliminary panel for future development of a targeted microbiome profiling method for forensic human identification.

Finally, in **Chapter 3**, "Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification" (Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, and Budowle B. 2017. *Forensic Sci Int Genet* (submitted)), describes the development of a novel targeted enrichment and sequencing method, the hidSkinPlex, for skin microbiome profiling for forensic human identification. The hidskinPlex is comprised of 286

clade-specific markers from 22 bacterial (and phage) clades at the family, genus, species, and subspecies level, which were selected and described in Chapter 2, as candidate markers to differentiate individuals based on their unique skin microbiome profiles. The performance of the hidSkinPlex was evaluated using bacterial control samples to assess the sensitivity and specificity of the panel, amplification (or coverage) and read depth of each marker, and uniformity of read depth across markers. To further evaluate the performance of the hidSkinPlex for prediction purposes, the hidSkinPlex was used to generate marker profiles from skin swab samples collected from eight individuals and three body sites (in triplicate). RMLR and 1NN were performed to attribute skin microbiome samples to their donor hosts. Classification was assessed for each body site and all samples together, regardless of body site. Classification accuracies calculated using enriched marker data were compared to accuracies generated in Chapter 2, using shotgun metagenomic data. Lastly, a case study was performed to compare human STR/SNP profiles generated from the collected skin swab samples to corresponding hidSkinPlex profiles to highlight the potential of using microbiome profiles independently or in conjunction with human forensic profiles for low-biomass samples.

The studies comprising this body of work provide a new tool, AutoCurE, for curating a local bacterial genome database for use in comparative genomic studies, a new method for identifying individual-specific skin microbiome signatures for application for human identification, and a novel targeted sequencing panel, the hidSkinPlex, to use for skin microbiome profiling for forensic human identification. Future studies will focus on evaluating the hidSkinPlex on larger sets of population samples, assessing the stability and diversity of skin microbiomes over time, performance of the panel on forensic samples, and assessment of analysis methods and interpretation guidelines for using the hidSkinPlex in a forensic setting.

REFERENCES

1. Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JAL, Marrone BL, Robertson J, Campos J. 2003. Building microbial forensics as a response to bioterrorism. Science 301:1852–1853.

2. Schmedes SE, Sajantila A, Budowle B. 2016. Expansion of Microbial Forensics. J Clin Microbiol 54:1964–1974.

3. Fraser-Liggett CM. 2005. Insights on biology and evolution from microbial genome sequencing. Genome Res 15:1603–1610.

4. Joint Genome Institute. 2017. Integrated Microbial Genomes & Microbiome Samples. https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=ImgStatsOverview

5. Human Microbiome Jumpstart Reference Strains Consortium. 2010. A catalog of reference genomes from the human microbiome. Science 328:994–999.

6. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. Nucleic Acids Res 43:D30–D35.

7. Amid C, Birney E, Bower L, Cerdeño-Tárraga A, Cheng Y, Cleland I, Faruque N, Gibson R, Goodgame N, Hunter C, Jang M, Leinonen R, Liu X, Oisel A, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Rivière S, Rossello M, Senf A, Smirnov D, Ten Hoopen P, Vaughan D, Vaughan R, Zalunin V, Cochrane G. 2012. Major submissions tool developments at the European nucleotide archive. Nucleic Acids Res 40:D43-47.

8. Kodama Y, Mashima J, Kaminuma E, Gojobori T, Ogasawara O, Takagi T, Okubo K, Nakamura Y. 2012. The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. Nucleic Acids Res 40:D38–D42.

9. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060.

10. Kyrpides NC, Woyke T, Eisen JA, Garrity G, Lilburn TG, Beck BJ, Whitman WB, Hugenholtz P, Klenk H-P. 2014. Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. Stand Genomic Sci 9:1278–1296.

11. Kyrpides NC, Hugenholtz P, Eisen J a, Woyke T, Göker M, Parker CT, Amann R, Beck BJ, Chain PSG, Chun J, Colwell RR, Danchin A, Dawyndt P, Dedeurwaerdere T, DeLong EF,

Detter JC, De Vos P, Donohue TJ, Dong X-Z, Ehrlich DS, Fraser C, Gibbs R, Gilbert J, Gilna P, Glöckner FO, Jansson JK, Keasling JD, Knight R, Labeda D, Lapidus A, Lee J-S, Li W-J, Ma J, Markowitz V, Moore ERB, Morrison M, Meyer F, Nelson KE, Ohkuma M, Ouzounis CA, Pace N, Parkhill J, Qin N, Rossello-Mora R, Sikorski J, Smith D, Sogin M, Stevens R, Stingl U, Suzuki K-I, Taylor D, Tiedje JM, Tindall B, Wagner M, Weinstock G, Weissenbach J, White O, Wang J, Zhang L, Zhou Y-G, Field D, Whitman WB, Garrity GM, Klenk H-P. 2014. Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. PLoS Biol 12:e1001920.

12. Whitman WB, Woyke T, Klenk H-P, Zhou Y, Lilburn TG, Beck BJ, De Vos P, Vandamme P, Eisen JA, Garrity G, Hugenholtz P, Kyrpides NC. 2015. Genomic Encyclopedia of Bacterial and Archaeal Type Strains, Phase III: the genomes of soil and plant-associated and newly described type strains. Stand Genomic Sci 10:26.

13. Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

14. Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature 486:215–221.

15. Gilbert JA, Jansson JK, Knight R. 2014. The Earth Microbiome project: successes and aspirations. BMC Biol 12:69.

16. Markowitz VM, Chen I-M a, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, Liolios K, Pagani I, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res 40:D123–D129.

17. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481.
18. Leake SL, Pagni M, Falquet L, Taroni F, Greub G. 2016. The salivary microbiome for differentiating individuals: proof of principle. Microbes Infect 1–7.

19. Franzosa E a., Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci 112:E2930–E2938.

20. Giampaoli S, Berti A, Valeriani F, Gianfranceschi G, Piccolella A, Buggiotti L, Rapone C, Valentini A, Ripani L, Romano Spica V. 2012. Molecular identification of vaginal fluid by microbial signature. Forensic Sci Int Genet 6:559–564.

21. Choi A, Shin K-J, Yang WI, Lee HY. 2014. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med 128:33–41.
22. Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S, Tomberlin JK. 2014. The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. Int J Legal Med 128:193–205.

23. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale J V., DeBruyn JM, Lents NH, Oh J, Byrd AL, Park M, Kong HH, Segre JA, Bashan A, Gibson T, Friedman J, Carey V, Weiss S, Hohmann E, Jortha P, Turner K, Gumus P, Nizam N, Buduneli N, Whiteley M, Hyde E, Haarmann D, Lynne A, Bucheli S, Petrosino J, Metcalf JL, Parfrey LW, Gonzalez A, Lauber CL, Knights D, Ackermann G, Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S, Tomberlin JK, Pechal JL, Crippen TL, Tarone AM, Lewis AJ, Tomberlin JK, Benbow ME, Carter DO, Metcalf JL, Bibat A, Knight R, Cobaugh KL, Schaeffer SM, DeBruyn JM, Finley SJ, Pechal JL, Benbow ME, Robertson BK, Javan GT, Hauther KA, Cobaugh KL, Jantz LM, Sparer TE, DeBruyn JM, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Guyer M, Michaud J-P, Gaétan M, Bishop C, Hill MO, Alpaydin E, Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Basak D, Srimanta P, Patranabis DC, Hoerl AE, Kennard RW, Liaw A, Matthew W, Saeys Y, Inza I, Larrañaga P, Guyon I, Elisseeff A, Metcalf JL, Xu ZZ, Weiss S, Lax S, Treuren W Van, Hyde ER. 2016. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS One 11:e0167370.

24. Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen J a, Gilbert J a. 2015. Forensic analysis of the microbiome of phones and shoes. Microbiome 3:21.

25. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. Nature 457:480–484.

26. Metzker ML, Mindell DP, Liu X-M, Ptak RG, Gibbs RA, Hillis DM. 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proc Natl Acad Sci U S A 99:14292–14297.

27. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. 2010. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proc Natl Acad Sci U S A 107:21242–21247.

28. González-Candelas F, Bracho MA, Wróbel B, Moya A. 2013. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. BMC Biol 11:76.

29. Grice EA, Segre JA. 2012. The Human Microbiome: Our Second Genome. Annu Rev Genomics Hum Genet 13:151–170.

30. Savage DC. 1977. Microbial Ecology of the Gastrointestinal Tract. Annu Rev Microbiol 31:107–133.

31. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. bioRxiv. doi: 10.1101/036103. http://biorxiv.org/content/early/2016/01/06/036103.abstract

32. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES.

2007. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104:19428–19433.

33. Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260–270.

34. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694–1697.

35. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

36. Fierer N, Hamady M, Lauber CL, Knight R. 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci U S A 105:17994–17999.

37. Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L. 2010. Short-term antibiotic treatment has differing long- term impacts on the human throat and gut microbiome. PLoS One 5:e9836.

38. Yassour M, Vatanen T, Siljander H, Hämäläinen A, Härkönen T, Ryhänen SJ, Franzosa EA, Vlamakis H. 2016. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med 8:343ra81.

39. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1031.

40. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, Goedert JJ, Hayes RB, Yang L. 2013. Human Gut Microbiome and Risk of Colorectal Cancer. J Natl Cancer Inst 105:1907–1911.

41. Kassinen A, Krogius-Kurikka L, Mäkivuokko H, Rinttilä T, Paulin L, Corander J, Malinen E, Apajalahti J, Palva A. 2007. The Fecal Microbiota of Irritable Bowel Syndrome Patients Differs Significantly From That of Healthy Subjects. Gastroenterology 133:24–33.

42. Tilg H. 2010. Obesity, Metabolic Syndrome, and Microbiota Multiple Interactions. J Clin Gastroenterol 44:16–18.

43. Lambert JA, John S, Sobe JD, Akins RA. 2013. Longitudinal analysis of vaginal microbiome dynamics in women with recurrent bacterial vaginosis: Recognition of the conversion process. PLoS One 8:e82599.

44. Califf K, Gonzalez A, Knight R, Caporaso JG. 2014. The Human Microbiome : Getting

Personal. Microbe 9:410–415.

45. Sommer MO, Dantas G, Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325:1128–1131.

46. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. 2013. Computational meta'omics for microbial community studies. Mol Syst Biol 9:666.

47. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642.

48. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072.

49. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

50. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761–2764.

51. Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625–630.

52. Soergel DAW, Dey N, Knight R, Brenner SE. 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J 6:1440–1444.

53. Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328–1333.

54. Wang Y, Zhang Z, Ramanan N. 1997. The actinomycete Thermobispora bispora contains two distinct types of transcriptionally active 16S rRNA genes. J Bacteriol 179:3270–3276.

55. Fox GE, Wisotzkey JD, Jurtshuk, Jr. P. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170.

56. Asai T, Zaporojets D, Squires C, Squires CL. 1999. An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc Natl Acad Sci U S A 96:1971–1976.

57. Schouls LM, Schot CS, Jacobs JA. 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 185:7241–7246.

58. Santiago-Rodriguez T, Cano R. 2016. Soil Microbial Forensics. Microbiol Spectr 1–15.

59. Hares DR. 2015. Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet 17:33–34.

60. Budowle B, Van Daal A. 2008. Forensically relevant SNP classes. Biotechniques 44:603–610.

61. Budowle B, Eisenberg AJ, van Daal A. 2009. Validity of low copy number typing and applications to forensic science. Croat Med J 50:207–217.

62. Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 108:68–74.

63. Holland MM, Parsons TJ. 1999. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Sci Rev. 11:21-50.

64. Davis C, Peters D, Warshauer D, King J, Budowle B. 2015. Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing. Leg Med 17:123–127.

65. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12C:128–135.

66. Grice E a, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre J a. 2008. A diversity profile of the human skin microbiota. Genome Res 18:1043–1050.

67. Quagliarello B, Cespedes C, Miller M, Toro A, Vavagiakis P, Klein RS, Lowy FD. 2002. Strains of Staphylococcus aureus obtained from drug-use networks are closely linked. Clin Infect Dis 35:671–677.

68. Grice E a, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre J a. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190–1192.

69. Capone KA, Dowd SE, Stamatas GN, Nikolovski J. 2011. Diversity of the human skin microbiome early in life. J Invest Dermatol 131:2026–2032.

70. Oh J, Conlan S, Polley EC, Segre JA, Kong HH. 2012. Shifts in human skin and nares microbiota of healthy children and adults. Genome Med 4:77.

71. Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, Gouilh MA, Pariente K, Segondy M, Burguière A, Manuguerra JC, Caro V, Eloit M. 2012. Human skin Microbiota: High

diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One 7:e38499.

72. Mathieu A, Delmont TO, Vogel TM, Robe P, Nalin R, Simonet P. 2013. Life on Human Surfaces: Skin Metagenomics. PLoS One 8:e65288.

73. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, NIH Intramural Sequencing Center Comparative Sequencing Program, Kong HH, Segre JA. 2013. Topographic diversity of fungal and bacterial communities in human skin. Nature 498:367–370.

74. Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA. 2014. Biogeography and individuality shape function in the human skin metagenome. Nature 514:59–64.

75. Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, Sanmiguel AJ, Minot S, Bushman FD, Grice EA, Grice A. 2015. The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. MBio 6:e01578-15.

76. Oh J, Byrd AL, Park M, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854–866.

77. Li K, Bihan M, Methé BA. 2013. Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome. PLoS One 8:e63139.

78. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC, Komarow HD, NISC Comparative Sequence Program, Murray PR, Turner ML, Segre JA. 2012. Temporal shifts in the skin microbiome associated with disease flare and treatment in children with atopic dermatitis. Genome Res 22:850–859.

79. Conlan S, Mijares LA, Becker J, Blakesley RW, Bouffard GG, Brooks S, Coleman H, Gupta J, Gurson N, Park M, Schmidt B, Thomas PJ, Otto M, Kong HH, Murray PR, Segre JA. 2012. Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome Biol 13:R64.

80. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vázquez-Baeza Y, Gonzalez A, Knight R, Dunn RR, Fierer N. 2014. Temporal variability is a personalized feature of the human microbiome. Genome Biol 15:531.

81. Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown G, Green JL, Bohannan BJ. 2015. Humans differ in their personal microbial cloud. PeerJ 3:e1258.

82. Goga H. 2012. Comparison of bacterial DNA profiles of footwear insoles and soles of feet for the forensic discrimination of footwear owners. Int J Legal Med 126:815–823.

83. Meadow JF, Altrichter AE, Green JL. 2014. Mobile phones carry the personal microbiome

of their owners. PeerJ 2:e447.

84. Williams DW, Gibson G. 2017. Individualization of pubic hair bacterial communities and the effects of storage time and temperature. Forensic Sci Int Genet 26:12–20.

85. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343–59.

86. Nishi E, Tashiro Y, Sakai K. 2014. Discrimination among individuals using terminal restriction fragment length polymorphism profiling of bacteria derived from forensic evidence. Int J Legal Med 129:425–433.

87. Nishi E, Watanabe K, Tashiro Y, Sakai K. 2017. Terminal restriction fragment length polymorphism profiling of bacterial flora derived from single human hair shafts can discriminate individuals. Leg Med 25:75–82.

88. Nayfach S, Pollard KS. 2015. Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria. bioRxiv DOI:10.1101/031757.

CHAPTER 1

Correcting Inconsistencies and Errors in Bacterial Genome Metadata Using an Automated Curation Tool in Excel (AutoCurE)

Published in Frontiers in Bioengineering and Biotechnology
2015, 3:138

Sarah E. Schmedes
Jonathan L. King
Bruce Budowle

ABSTRACT

Whole-genome data are invaluable for large-scale comparative genomic studies. Current sequencing technologies have made it feasible to sequence entire bacterial genomes with relative ease and time with a substantially reduced cost per nucleotide, hence cost per genome. More than 3,000 bacterial genomes have been sequenced and are available at the finished status. Publically available genomes can be readily downloaded; however, there are challenges to verify the specific supporting data contained within the download and to identify errors and inconsistencies that may be present within the organizational data content and metadata. AutoCurE, an automated tool for bacterial genome database curation in Excel, was developed to facilitate local database curation of supporting data that accompany downloaded genomes from the National Center for Biotechnology Information. AutoCurE provides an automated approach to curate local genomic databases by flagging inconsistencies or errors by comparing the downloaded supporting data to the genome reports to verify genome name, RefSeq accession numbers, the presence of archaea, BioProject/UIDs, and sequence file descriptions. Flags are generated for nine metadata fields if there are inconsistencies between the downloaded genomes and genomes reports and if erroneous or missing data are evident. AutoCurE is an easy-to-use tool for local database curation for large-scale genome data prior to downstream analyses.

KEYWORDS: Bacteria · Genomes · Metadata · Database · Curation · Automation · AutoCurE

INTRODUCTION

Advancements in sequencing technologies in the past several years have resulted in a substantial increase in the number of bacterial genomes that have been and continue to be sequenced. The first complete bacterial genome was sequenced in 1995 (Fleischmann et al., 1995) and 24 microbial organisms were completely sequenced within the next 5 years (Nierman et al., 2000). Ten years later, in 2005, there were almost 300 prokaryote genomes sequenced (Fraser-Liggett, 2005) and as of May 2015 there were 34,066 bacterial genomes available at the complete (3,725), chromosome (773), scaffold (11,028), and contig (18,540) status as listed by the National Center for Biotechnology Information (NCBI)[1]. Integrated Microbial Genomes (IMG)[2] (Markowitz et al., 2012) reported the number of bacterial genomes at 26,033 at the finished (3,378), draft (1,683), and permanent draft (20,972) status, and there is a total of 39,969 bacterial genome sequencing projects listed in the Genomes OnLine Database (GOLD)[3] (Reddy et al., 2015), an increase from only 1,986 in 2007. As a result of advancements in sequencing technologies, with increased output and decreased costs, the number of completed genomes will continue to rise resulting in substantial amounts of data.

These whole bacterial genome sequence data are housed in publically available databases such as NCBI[4] (Benson et al., 2015), European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI)[5] (Amid et al., 2012), and DNA Data Bank of Japan (DDBJ)[6] (Kodama et al., 2012), which make up the International Nucleotide Sequence Database Collaboration (INSDC) (Nakamura et al., 2013). Additional databases with more specific

[1] http://www.ncbi.nlm.nih.gov/genome/browse/, accessed May 28, 2015

[2] https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=ImgStatsOverview, accessed May 28, 2015

[3] https://gold.jgi-psf.org/statistics

[4] http://www.ncbi.nlm.nih.gov/genbank/

[5] http://www.ebi.ac.uk/ena

[6] http://www.ddbj.nig.ac.jp/

microbial applications and bioin-formatics programs include IMG (Markowitz et al., 2012) and PATRIC (Pathosystems Resource Integration Center) (Wattam et al., 2014). Data can be readily downloaded from these data-bases through ftp sites or facilitated through download links. The NCBI ftp site[7] provides links to download all bacterial genomes in a number of file types. However, since these downloads include thousands of complete bacterial genomes, there is a challenge to easily identify which genomes were included in the download, to determine if all files and metadata associated with particular genomes were included and whether supporting data were correct. Quality control of supporting data within public databases is crucial to ensure accurate and the most up-to-date metadata; however, quality control practices and methods are not readily known or clearly stated. Inaccurate identifying information can confound downstream analyses and may cause misinterpretations and therefore curation of metadata is necessary. High-quality databases are essential for research areas, such as comparative genomics, phylogenetics, and metagenomics, especially as they apply to diagnostics, public health, biosafety and biosecurity, and microbial forensics.

In this study, a local database was created that contained all publically available complete bacterial genomes from the NCBI ftp site. Metadata inconsistencies were observed between the downloaded genomes and those listed as complete genomes on the genome reports from NCBI Genome. To use these data for downstream studies, a manual curation was performed to identify and correct inconsistencies and to delete erroneous files. Manual curation was performed to compare the supporting data associated with each sequence file, including genome name, UID (unique identifier) number, RefSeq accession numbers, and file descriptions found within each file to the metadata included in the complete genome reports. The process was performed using a "one-

[7] ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/

by-one" approach which was time consuming and not routinely practical for future efforts, especially as the number of genome entries continues to increase. Therefore, an automated tool for bacterial database curation in Excel (AutoCurE)[8] was developed, decreasing the curation time from months with manual curation to minutes with automated curation. AutoCurE facilitates checks between the downloaded genome folders, files and the genome reports to flag if any inconsistencies exist in the metadata, including genome names, BioProject/UID, RefSeq accession numbers, and sequence file descriptions, and to identify and flag archaea genomes.

MATERIALS AND METHODS

Genomes

All complete bacterial genomes were downloaded on March 5, 2014 from the Bacteria folder on the NCBI ftp site[7] using the all.fna.tar.gz link to retrieve all fna files (DNA genome sequence in FASTA format). Genome reports of all complete bacteria and archaea genomes were downloaded from NCBI Genome[9] on March 6, 2014. No modification dates were listed on the genome reports for March 5, 2014 to March 6, 2014 (to rule out discrepancies between the genome file and genome report download dates).

Manual Curation

Manual curation of the local bacterial genomes database was performed in three rounds. In Round 1, downloaded genome folder names were compared with the complete bacterial and archaeal genome reports to identify archaea genomes and bacterial genomes found in the genome

[8] AutoCurE is available and maintained by the Institute of Applied Genetics at https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular- and-medical-genetics/laboratory-faculty-and-staff/AutoCurE/. AutoCurE will be updated based on user feedback and any changes made to genome report formats and/or structure by NCBI.

[9] http://www.ncbi.nlm.nih.gov/genome/

report by name. In Round 2, genomes not found on either report were searched by RefSeq accession numbers from the files to identify genomes on reports that had been renamed. Any remaining genomes still not found on the genome reports were searched on NCBI to determine if the file had been discontinued or to verify the identity of the genome. In Round 3, "one-by-one" manual curation was performed to check genome names and files against the genome reports at the time of download in addition to current information on NCBI for genome name, BioProject/UID, file description, and RefSeq accession numbers.

AutoCurE Development and Features

AutoCurE was developed to provide an automated approach for bacterial database curation of downloaded supporting data from fna file types from the NCBI ftp site. AutoCurE is composed of two customized Excel workbooks, the AutoCurE Genome Filename Tool and the AutoCurE Genome Report Tool, with custom scripts and macros to: facilitate creating a print directory and file path of all downloaded genomes; rename all file names to the first line of text (to make the files more recognizable as opposed to just providing the accession number); parse out metadata fields to facilitate searches; and create flags to mark inconsistencies or errors between the downloaded genome files and the current bacteria and archaea genome reports. Flags are generated for nine different metadata categories to identify inconsistencies or errors pertaining to the following: (1 and 2) genome name for genus and species (strain was not included due to the wide variation of naming inconsistency of strain names); (3) to identify archaea; (4) to verify consistency between the original filename accession number and the accession number found within each sequence file; (5) to identify inconsistencies between the UID number from the genome folder name and the BioProject ID within the genome report; (6) identify if the RefSeq

accession number from each sequence file is found within the genome report; (7) identify accession numbers other than RefSeq reference assembly accessions (i.e., other than NC_ XXXXXX); (8) identify genome folders missing whole genome or chromosome files (i.e., only contains plasmid files); and (9) identify sequence files which may be a draft sequence. Report statements are generated for each flag to notify the user of potential changes or corrections that need to be made. AutoCurE also facilitates file manipulation by allowing the user to select sets of specific genomes and copies of the files are moved to a new directory to maintain an unaltered master copy of the database. This feature eliminates having to manually search and retrieve files for downstream use. All processing times reported were using a computer with i7-2600 CPU @ 3.4 GHz, 3.23 GB of RAM.

RESULTS

Manual Curation

All complete genomes in the Bacteria folder on the NCBI ftp site (*N* = 2,769; downloaded March 5, 2014) were downloaded to create a local bacterial genome database. Genome names from each of the genome folders were compared with the genome reports to separate bacteria genomes from archaea genomes (Table 1). Archaea genomes (*N* = 164) were found within the Bacteria folder and were removed. In addition, 157 genomes were not found on either report by genome name. In order to verify the identity of these genomes, RefSeq accession numbers (as listed as the sequence file names) were searched against each genome report. Of these genomes searched, 87 bacterial genomes were found on the genome report associated with a different genome name; 59 bacterial genomes and 5 archaeal genomes were not found on the genome report but were found

in the NCBI Nucleotide database; and 6 bacterial genomes had been removed by NCBI, and the accession numbers had been discontinued.

Table 1. Inconsistencies between genome downloads and genome reports.

		Bacteria	Archaea	Total
Round 1 (genome name search)	Downloaded genomes from ftp site	2,605	164	2,769
	Complete genomes listed in genome report	2,734	168	2,902
	Downloaded genome names found within report	2,453	159[a]	2,612
	Not found in genome report by name			157
Round 2 (genome accession number search)	Accession number found in genome report, genome name change (includes strain name)	87		
	Accession numbers discontinued	6		
	Accession number not found in genome report but found on NCBI Nucleotide	59	5	64
Round 3 ("one- by-one" manual curation)	Starting number of genomes			2,599
	No inconsistencies or errors observed	2,402		
	Not found on genome report	57		
	Found on genome report but no accession numbers listed	18		
	Genus and/or species name inconsistent	68		
	Potential draft sequence	56		
	Chromosome/genome data missing (only plasmid files present)	5		
	Changed from complete status	131		
	Genome folder contained erroneous files	9		
	Genomes deleted for not containing complete reference assemblies	19		
	Final number of bacterial genomes in local database			2,580

[a]*Thermoproteus tenax Kra 1 was found in both the bacteria and archaea genome reports.*

Although the majority of the genome folders (N = 2,402) were named correctly and contained the correct files, other types of errors and inconsistencies were observed. These problematic data included duplicate genome names, non-reference assembly file types (i.e., contig or scaffold files, environmental sequence files, and genome folders only containing plasmids), naming inconsistencies, and files misplaced in genome folders. In order to verify all files and associated metadata, a "one- by-one" manual curation was performed on 2,599 bacterial genomes in the database after all archaea and discontinued genomes had been removed. Downloaded genome folder names, BioProject/UID numbers, and sequence file descriptions and accession

numbers (first line of text within fna files) were compared with the metadata included in the genome reports to identify inconsistencies. The most common discrepancies were inconsistent genome name nomenclature between the genome folder name, genome report, and within the fna file (*N* = 68; genus and species names), indicating inconsistent updates when genome names are changed. For example, Candidatus *Endolissoclinum faulkneri* L2, BioProject PRJNA182483, had a genome folder named *Thalassobaculum* L2 containing an fna file with the genome name Candidatus *Endolissoclinum patella* L2; thus, illustrating the discrepancies that can occur between the genome report, genome folder, and fna file. Inconsistent genome names also included likely major spelling errors. In addition, there were examples of genome folders having the same name, including strain (*N* = 83), with the only differences in the accession numbers and BioProject IDs. The bacteria genome report contained 126 duplicate genome names, 7 of which had duplicate BioProject IDs, and 64 of the duplicate genome names were associated with the 83 duplicate genome folder names. While duplicate genome names were not considered errors, as these are the correct names, it does point out the need for better naming requirements (such as substrain or isolate ID) to differentiate another genome from another in addition to solely the BioProject ID. In addition, 57 genomes were not found on the genome report but were found on NCBI Nucleotide, and 18 genomes were found on the report but had no accession numbers associated with the genome. Additionally, 19 genome folders were removed due to RefSeq accession numbers for associated fna files being discontinued and removed from NCBI, RefSeq accession numbers not listed on genome reports or genome page, genome status changed to scaffold-level, genome folder only contained plasmid files, and not all chromosome files were included in genome folder, resulting in 2,580 genomes in the local database. Round 3 manual curation includes the results found within Round 2, and the results are more inclusive.

Erroneous files were also found within the downloaded data and associated with incorrect genomes. The downloaded data were retrieved from the complete bacterial genomes folder on the NCBI ftp site; however, 56 genomes were found as potential draft genomes (i.e., text within fna files listed these sequences as “draft,” “partial sequence,” “provisional sequence,” “nearly complete genome,” “sequencing in progress,” and “non-contiguous finished genome”), 5 genome folders only contained plasmid files, and in the course of manual curation, 3 genomes were changed to scaffold status and 128 genomes changed from “complete” to “chromosome” or “chromosome with gaps” status. In addition, nine genome folders contained erroneous files which either did not belong to that particular genome or were not RefSeq reference assembly files. For example, the genome folder for *Vibrio parahaemolyticus* O1 K33 CDC K4557 contained the two correct chromosome files for this genome; however, an additional 17 files were found within the genome folder belonging to a different strain of *Vibrio parahaemolyticus*, 9 different strains of *Listeria monocytogenes*, and 1 strain of *Campylobacter jejuni*. Additionally, at the time of download, there were six different substrains of *Synechocystis* sp. PCC 6803, of which three substrains, including GT-I, PCC-N, and PCC-P, had incorrectly associated substrain names, BioProject/UIDs, and fna files.

AutoCurE

AutoCurE was developed to automate curation of supporting data of local bacterial genome databases from data downloaded from the NCBI ftp site. AutoCurE is composed of two Excel work-books, the AutoCurE Genome Filename Tool and the AutoCurE Genome Report Tool, with customized scripts to automatically generate flags for nine different metadata categories to identify inconsistencies and errors of the types found during manual curation, which are listed in Table 2.

After whole-genome data and genome reports are downloaded, AutoCurE generates print lists of the genomes downloaded and compares this list to the bacteria and archaea genome reports to identify archaea and compare the genome name, BioProject/UIDs, RefSeq accession numbers, and fna sequence file descriptions to flag inconsistencies between the downloaded data and genome reports. Report statements are generated for each flag, notifying the user of corrections that may need to be made to the local database.

Table 2. AutoCurE genome filename and report tools.

Features
Prints list directory of downloaded genomes and file paths
Pulls out first line of text from files to provide RefSeq accession number and sequence file description
Parses metadata from genome reports and data downloads into lists to compare BioProject/UID, RefSeq accession number, genome folder name, file name, and file description
File manipulation to eliminate manual searching within directories. Allows the user to check desired genomes in the Excel workbook and a copy of the genome files is made to another directory for downstream use, thus keeping an unaltered master copy of the database
Flags
Accession number in genome report: indicates if the accession number within the sequence file is found in the genome report
BioProject/UID match: compares the UID from downloaded genome folder names to BioProject ID in the genome report
Original accession consistency: indicates if the original accession number file name matches the accession number found within the sequence file
Genus match: compares genus name from genome report to genome folder to sequence file
Species match: compares species name from genome report to genome folder to sequence file
Archaea: identifies archaea genomes based on accession numbers found in the archaea genome report
RefSeq reference genome accession: identifies any files with an accession number other than a RefSeq reference assembly number
Chromosome/genome data present: indicates if only plasmid sequence files are present (i.e., chromosome or whole-genome data are absent)
Draft or partial sequence: identifies any potential draft genome or partial sequences based on sequence file description

The same genome dataset from Round 2 manual curation was used to validate the ability of AutoCurE to compare the automated results with the manual curation results. In addition, 10 archaea genomes were included to validate the archaea flag, since all known archaea found on the archaea genome report had been removed prior to the Round 2 genomes dataset. AutoCurE processed 2,621 genomes (4,956 files) in less than 30 min. In comparison, manual curation took several months (with a 2–3 days per week effort). By default, AutoCurE can process up to 10,000 files; however, more genomes/files can be easily accommodated with minor formula modifications. Flags were successfully generated for each of the nine categories. Figure 1 shows an example of the AutoCurE Genome Report Tool flagging multiple fna files in the accession number and genus and species name categories. Report statements were generated for each flag, indicating potential changes which need to be made to the database files or metadata. Each flag was manually checked to ensure that flags were appropriately generated.

	E	F	I	J	K	L	M	N
2		0	Copy Genome Folder	#Organism/Name from Genome Report	Genome Name from Folder	UID from Folder	Accession Number from FLT	Description from FLT
616		257		Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis KBAB4	uid58315	NC_010180.1	Bacillus weihenstephanensis KBAB4 plasmid pBWB401 ,CS.fna
617		257		Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis KBAB4	uid58315	NC_010181.1	Bacillus weihenstephanensis KBAB4 plasmid pBWB402 ,CS.fna
618		257		Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis KBAB4	uid58315	NC_010182.1	Bacillus weihenstephanensis KBAB4 plasmid pBWB403 ,CS.fna
619		257		Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis KBAB4	uid58315	NC_010183.1	Bacillus weihenstephanensis KBAB4 plasmid pBWB404 ,CS.fna
620		257		Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis KBAB4	uid58315	NC_010184.1	Bacillus weihenstephanensis KBAB4 chromosome ,CG.fna
621		258		Bacteriovorax marinus SJ	Bacteriovorax marinus SJ	uid82341	NC_016620.1	Bacteriovorax marinus SJ ,CG.fna
622		258		Bacteriovorax marinus SJ	Bacteriovorax marinus SJ	uid82341	NC_019100.1	Bacteriovorax marinus SJ plasmid pBMS1 ,CS.fna
623		259		Liberibacter crescens BT 1	bacterium BT 1	uid184079	NC_019907.1	Liberibacter crescens BT 1 chromosome ,CG.fna
624		260		#N/A	Bacteroides CF50	uid222805	NC_022526.1	Bacteroides sp. CF50 ,CG.fna
625		261		Bacteroides fragilis 638R	Bacteroides fragilis 638R	uid84217	NC_016776.1	Bacteroides fragilis 638R ,CG.fna
626		262		Bacteroides fragilis NCTC 9343	Bacteroides fragilis NCTC 9343	uid57639	NC_003228.3	Bacteroides fragilis NCTC 9343 ,CG.fna
627		262		Bacteroides fragilis NCTC 9343	Bacteroides fragilis NCTC 9343	uid57639	NC_006873.1	Bacteroides fragilis NCTC 9343 plasmid pBF9343 ,CS.fna
628		263		Bacteroides fragilis YCH46	Bacteroides fragilis YCH46	uid58195	NC_006297.1	Bacteroides fragilis YCH46 plasmid pBFY46 ,CS.fna
629		263		Bacteroides fragilis YCH46	Bacteroides fragilis YCH46	uid58195	NC_006347.1	Bacteroides fragilis YCH46 DNA ,CG.fna

	N	O	P	Q	S	T	U	V	W	X	Y
2	Description from FLT	Accession Number in Genome Report	BioProject/UID Match	Original Accession Consistency	Genus Match	Species Match	Archaea	RefSeq Reference Genome Accession	Chromosome/ Genome Data Present	Draft or Partial Sequence	Report Statement
616	Bacillus weihenstephanensis KBAB4 plasmid pBWB401 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
617	Bacillus weihenstephanensis KBAB4 plasmid pBWB402 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
618	Bacillus weihenstephanensis KBAB4 plasmid pBWB403 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
619	Bacillus weihenstephanensis KBAB4 plasmid pBWB404 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
620	Bacillus weihenstephanensis KBAB4 chromosome ,CG.fna	✓	✓	✓	✓	✓		✓	✓		
621	Bacteriovorax marinus SJ ,CG.fna	✓	✓	✓	✓	✓		✓	✓		
622	Bacteriovorax marinus SJ plasmid pBMS1 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
623	Liberibacter crescens BT 1 chromosome ,CG.fna	✓	✓	✓	✗	✗		✓	✓		The genus names listed in the geno
624	Bacteroides sp. CF50 ,CG.fna	✗		✓				✓	✓		Accession number NC_022526.1 wa
625	Bacteroides fragilis 638R ,CG.fna	✓	✓	✓	✓	✓		✓	✓		
626	Bacteroides fragilis NCTC 9343 ,CG.fna	✓	✓	✓	✓	✓		✓	✓		
627	Bacteroides fragilis NCTC 9343 plasmid pBF9343 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
628	Bacteroides fragilis YCH46 plasmid pBFY46 ,CS.fna	✓	✓	✓	✓	✓		✓	✓		
629	Bacteroides fragilis YCH46 DNA ,CG.fna	✓	✓	✓	✓	✓		✓	✓		

Figure 1. AutoCurE genome report tool. AutoCurE compared content from the genome report, genome folder name, and fna file description to flag inconsistencies for nine metadata categories. Flags, shown as red Xs, were generated, indicating that a RefSeq accession number was not found in the genome report and inconsistencies in genus and species name. Additional columns in the Genome Report Tool, not shown, include a Comments section and metadata taken from the NCBI genome reports associated with each downloaded file. Columns E and F group the files associated with a particular genome by color (Column E) and by number (Column F). (FLT, First Line of Text within the fna file).

The total number of flags produced for each category by AutoCurE was consistent with values from manual curation with a few discrepancies. Discrepancies between the manually curated dataset and the AutoCurE dataset include genome name inconsistencies in the species

name category when a species name is not listed or when inconsistent punctuation may be present. A number of genus and species inconsistencies were also observed due to minor spelling errors (one letter difference). Since flags are generated based on customized formulas, anything outside the search parameter may be missed. For example, *Fibrella aestuarina* BUZ 2 genome was not flagged as a potential draft sequence due to a spelling error in the sequence file description, "drat genome." In addition, since *Thermoproteus tenax* Kra 1 was listed on both the bacteria and archaea genome reports but only had an accession number listed in the bacteria genome report, AutoCurE did not mark this genome as archaea, due to this error. Any errors within the genome report will not be flagged; only inconsistencies between the downloaded data and genome reports will be identified. Additionally, the File Management Center within the Genome Report Tool, which incorporates the file manipulation feature, was validated. More than 4,000 files were copied and moved to an output directory in less than 15 min. Smaller file batches (N = 50) can be moved in about 1 sec.

DISCUSSION

Whole-genome data are available at a number of public repositories. Some of these data are not necessarily curated, constantly being updated, and in flux. Therefore, it is expected that errors and inconsistencies will arise, such as in genome names, since taxonomic name changes occur frequently. One should be aware of the types of inconsistencies and errors that are and may be present in order to correct them before using the data for research and development in diagnostics, public health, biosafety and biosecurity, and microbial forensics. In this study, inconsistencies and errors were observed while creating a local bacterial genome database using whole-genome data available from the NCBI ftp site. The main issues observed included: archaea

genomes were colocalized in the same folder as the bacteria genomes; genome naming inconsistencies were observed between the genome folders, genome reports, and fna files; not all data downloaded were included in the genome reports and not all genomes found on the genome reports were available for download on the ftp site; discontinued files had not been removed from the ftp site; some genome folders contained draft genome or only plasmid files; and files were associated with incorrect genomes. In addition, during the course of manual curation, more than 130 genomes had been changed from "complete" to "chromosome," "chromosome with gaps," or "scaffold" status, indicating fluidity in genome status as genomes are updated; because of this lack of consistency, official genome status should be checked in GOLD (Reddy et al., 2015). Due to discrepancies in downloaded data from genome databases, proper curation is necessary prior to downstream use to reduce misinterpretations that may affect subsequent analyses.

As the number of available genomes continues to increase, it will not be practical to manually curate data. To reduce errors that may impact subsequent analyses, it is imperative to curate the downloaded data contained within local databases to remove redundancies, erroneous files, and correct for naming inconsistencies. An automated tool was needed to authenticate supporting data associated with downloaded publically available bacterial genomes. AutoCurE was developed to facilitate curation of local bacterial databases by reducing curation time from months to minutes while automatically flagging errors and inconsistencies. Other tools have been developed for local database storage and manipulation, such as MicrobeDB (Langille et al., 2012). MicrobeDB is a Linux-based database tool facilitating genome downloads from the NCBI ftp site, archiving files and updating the database, and database manipulation (Langille et al., 2012). While a useful tool for bacterial database manipulation, MicrobeDB requires the user to be familiar with Perl programming or with MySQL (Langille et al., 2012). In contrast, AutoCurE is Excel- based

to provide ease-of-use in a Windows-based platform for metadata curation and database manipulation, which may be better suited for users not adept at programming.

There is a need for better quality checks as databases are maintained and updated to check for naming inconsistencies/ changes, updating sequence files, removing discontinued files, and checking for correct file placement and UID associations. Recommendations and changes have been made by the INSDC to replace genome identifiers from strain taxids with alternative, more unique metadata, such as BioSample, BioProject, or assembly ID (Federhen et al., 2014). Moving toward a metadata system of more unique identifiers helps reduce ambiguities when genomes are named with the same strain name. However, improved quality control of database management needs to be implemented to maintain the most up-to-date and accurate files and metadata on public repository sites. AutoCurE provides a solution for automated curation of these supporting data to provide a quality check prior to using the downloaded files, thus eliminating the need for manual curation or downloading each genome one at a time. Improved upfront quality control of data directly by public database managers would reduce the need for downstream resources and provide a seamless flow of higher quality data and metadata directly to the end user. As genome data continue to grow, quality control practices and additional tools, such as AutoCurE, are exceedingly important for data storage, curation, and manipulation.

AUTHOR CONTRIBUTIONS

SS and BB conceived the project. SS and JK designed and developed the analytical tools. SS analyzed the data. SS, JK, and BB wrote the article.

FUNDING

This project was supported by internal funds from the Institute of Applied Genetics, University of North Texas Health Science Center.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Amid, C., Birney, E., Bower, L., Cerdeño-Tárraga, A., Cheng, Y., Cleland, I., et al. (2012). Major submissions tool developments at the European nucleotide archive. *Nucleic Acids Res.* 40, D43–D47. doi:10.1093/nar/ gkr946

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2015). GenBank. *Nucleic Acids Res.* 43, D30–D35. doi:10.1093/nar/ gku1216

Federhen, S., Clark, K., Barrett, T., Parkinson, H., Ostell, J., Kodama, Y., et al. (2014). Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and assembly records. *Stand. Genomic Sci*. 9, 1275–1277. doi:10.4056/sigs.4851102

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512. doi:10.1126/ science.7542800

Fraser-Liggett, C. M. (2005). Insights on biology and evolution from microbial genome sequencing. *Genome Res.* 15, 1603–1610. doi:10.1101/gr.3724205

Kodama, Y., Mashima, J., Kaminuma, E., Gojobori, T., Ogasawara, O., Takagi, T., et al. (2012). The DNA Data Bank of Japan launches a new resource, the DDBJ omics archive of functional genomics experiments. *Nucleic Acids Res.* 40, D38–D42. doi:10.1093/nar/gkr994

Langille, M. G. I., Laird, M. R., Hsiao, W. W. L., Chiu, T. A., Eisen, J. A., and Brinkman, F. S. L. (2012). MicrobeDB: a locally maintainable database of microbial genomic sequences. *Bioinformatics* 28, 1947–1948. doi:10.1093/ bioinformatics/bts273

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi:10.1093/nar/gkr1044

Nakamura, Y., Cochrane, G., and Karsch-Mizrachi, I. (2013). The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 41, D21–D24. doi:10.1093/nar/gks1084

Nierman, W., Eisen, J. A., and Fraser, C. M. (2000). Microbial genome sequencing 2000: new insights into physiology, evolution and expression analysis. *Res. Microbiol.* 151, 79–84. doi:10.1016/S0923-2508(00)00125-X

Reddy, T. B. K., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., et al. (2015). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* 43, D1099–D1106. doi:10.1093/nar/gku950

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi:10.1093/nar/gkt1099

CHAPTER 2

Forensic Human Identification Using Skin Microbiomes

Published in Applied and Environmental Microbiology
2017 (In press)

Sarah E. Schmedes
August E. Woerner
Bruce Budowle

ABSTRACT

The human microbiome contributes significantly to the genetic content of the human body. Genetic and environmental factors help shape the microbiome, and as such, the microbiome can be unique to an individual. Previous studies have demonstrated the potential to use microbiome profiling for forensic applications, however a method has yet to identify stable features of skin microbiomes that produce high classification accuracies for samples collected over reasonably long time intervals. A novel approach is described to classify skin microbiomes to their donors by comparing two features types, *Propionibacterium acnes* pangenome presence/absence features and nucleotide diversities of stable clade-specific markers. Supervised learning was used to attribute skin microbiomes from 14 skin body sites from 12 healthy individuals sampled at three time points over a >2.5 year period with accuracies up to 100% for three body sites. Feature selection identified a reduced subset of markers from each body site that are highly individualizing, identifying 187 markers from 12 clades. Classification accuracies were compared in a formal model testing framework, and the results of this indicate that learners trained on nucleotide diversity perform significantly better than those trained on presence/absence encodings. This study used supervised learning to identify individuals with high accuracy and associated stable features from skin microbiomes over a period of up to almost 3 years. These selected features provide a preliminary marker panel for future development of a robust and reproducible method for skin microbiome profiling for forensic human identification.

KEYWORDS: Skin microbiome · Human identification · Forensic profiling · Metagenomics · Supervised learning

IMPORTANCE

A novel approach is described to attribute skin microbiomes, collected over a period of >2.5 years, to their individual hosts with a high degree of accuracy. Nucleotide diversities of stable clade-specific markers with supervised learning was used to classify skin microbiomes from a particular individual with up to 100% classification accuracy for three body sites. Attribute selection was used to identify 187 genetic markers from 12 clades which provide the greatest differentiation of individual skin microbiomes from 14 skin sites. This study performs skin microbiome profiling from a supervised learning approach and obtains high classification accuracy for samples collected from individuals over a relatively long time period for potential application to forensic human identification.

INTRODUCTION

The human microbiome plays a critical role in health, metabolism, and immune response (1) and can be influenced by numerous factors, including but not limited to genetics, geography, diet, and hygiene (2–4). Colonization of the human microbiome begins at birth and continues to change throughout development (5, 6), contributing an additional 5,000,000 genes from the gut microbiome alone (7) to the repertoire of human genes. Since unique genetic and environmental factors help shape the microbiome, the composition of the microbiome has the potential to be unique to its host individual. Features of the personal microbiome, such as strain-specific signatures (8, 9), which may be stable over time, make microbiome characterization potentially applicable to forensic human identification.

Current forensic human profiling methods typically utilize autosomal short tandem repeats (STRs) profiles to attribute forensic biological evidence to a suspect (or victim) (10). Often

evidentiary samples contain mixtures of human DNA from multiple sources or contain low amounts (i.e., low-copy number (LCN)) or degraded DNA, making interpretation of mixed or partial profiles difficult or inconclusive. In these cases alternative methods may be employed, such as sequencing high-copy number markers (e.g., targeting the hypervariable regions of the mitochondrial genome (11, 12) or whole mitochondrial genomes (13)), or methods to enhance sensitivity of detection including concentrating DNA extracts, increasing polymerase chain reaction (PCR) cycles, or performing whole-genome amplification (14). The human microbiome is an example of another high-copy number genetic marker, since microbial cells may be at a ratio of 1:1 (15) to 10:1 to human cells (16), and thus it is a potential target to complement partial or inconclusive STR profiles to increase resolution for human source attribution.

Recent studies have demonstrated the potential to use microbiome profiling for forensic identification, mainly using unsupervised methods to show that microbiome samples from touched objects resemble their respective donors (17–19). Few studies have addressed microbiome profiling from a supervised approach, i.e., for the purposes of classification. Franzosa et al. (8) used a nearest-neighbor classification approach using clade-specific markers and 1kb genomic windows to identify strain-level metagenomic codes specific to individuals; however this method could identify only <30% of individuals using skin microbiomes (i.e., anterior nares) sampled over a time interval of 30-300 days. Lax et al. (20) and Williams et al. (21) used random forests trained on operational taxonomic units (OTUs) abundances of targeted 16S rRNA sequences for human identification. While both approaches were highly accurate (96.3% and 97.3%, respectively), the samples were collected over short time intervals (< 3 days or just a single time point, respectively) (20, 21), making their results less applicable to a typical forensics setting.

Individual-specific microbiome features with the greatest temporal stability (up to almost 3 years) include single-nucleotide variant (SNV) profiles of *Propionibacterium acnes* from the skin (9) and gene signatures (i.e., clade-specific markers and 1kb genomic windows) from the gut microbiome (8). Strain-level signatures from shotgun sequencing provide far more depth of resolution than 16S rRNA based features, such as terminal restriction fragment length polymorphism profiles (18, 22, 23), OTUs abundances (8, 19–21, 24, 25), and biological community distances (e.g., UniFrac distance) (17, 20). Nucleotide diversity of strains, which measures the strain-level heterogeneity of the microbial population, also has been shown to be greater between individuals than within the same individual (26). Thus far, features used for microbiome profiling at the strain-level demonstrate the most success to differentiate individuals over time. However, a method has yet to be described that identifies differentiating features stable over reasonably long time intervals and applies appropriate measures on these markers to perform classification (i.e., via supervised learning) to attribute skin microbiome samples to their donors.

In this study, a novel approach is described to attribute skin microbiomes to their individual hosts with a high degree of accuracy and to identify genetic markers which may be well-suited to individual skin microbiome differentiation. Unsupervised learning techniques were first evaluated to assess inter- versus intra-sample variation across host microbiomes sampled across 14 body sites. To assess if microbiomes could be used to be predictive of their host, two feature types capturing strain-level variation within shotgun metagenomes were compared using two supervised learning techniques. In particular, *Propionibacterium acnes* pangenome presence/absence features and the nucleotide diversities of clade-specific markers were used in conjunction with regularized multinomial logistic regression (RMLR) and 1-nearest-neighbor (1NN) classifiers to form predictions on host microbiomes based on samples separated by up to three years. Feature selection

was then used to identify stable features which can be used to attribute skin microbiomes from multiple body sites to their respective hosts. This reduced set of markers was then evaluated to see if they could provide similar predictive power despite using much less information. The results from our classification algorithms were then formally compared to evaluate if different body sites and different classification techniques significantly vary in their predictive capabilities.

RESULTS

Sample and Shotgun metagenomic processing

Publically-available shotgun metagenomic datasets from Oh et al. (9) were used in this study. Briefly, the Oh et al. (9) dataset consists of an extensive spatial and temporal sampling of skin microbiomes from 12 healthy individuals across 17 body sites (i.e., antecubital fossa (Ac), alar crease (Al), back (Ba), cheek (Ch), external auditory canal (Ea), forehead (Fh), hypothenar palm (Hp), inguinal crease (Ic), interdigital web (Id), manubrium (Mb), occiput (Oc), popliteal fossa (Pc), plantar heel (Ph), retroauricular crease (Ra), toenail (Tn), and toe web space (Tw), and volar forearm (Vf)). Skin microbiome samples were collected at three different time points over a period of almost 3 years, sampled over long (ranging from 10-30 months) and short (ranging from 5-10 weeks) time intervals (9). In total, 2,446 fastq files from 585 samples, containing a total of 23 billion reads (mean of 39.3 million reads per sample) were downloaded from the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) (27). Data were pre-processed to remove sequencing adapters, trim reads with a quality score less than 20, remove reads less than 50 bp in length, and remove any human host-associated reads. A total of 12.6 billion quality-controlled reads (mean 21.5 million reads per sample) remained after read pre-processing. Several samples had substantially lower read depth after read pre-processing, and only individuals

with samples from all three time points at a particular body site, with ≥10x average read depth across all shared markers, were included in the study (n=381; Table S1). Three body sites from the foot (i.e., plantar heel (Ph), toenail (Tn), and toe web space (Tw)) also were excluded from the study, as they only shared 2-5 markers among samples.

Taxonomic classification was performed using MetaPhlAn2 (28) to identify the core skin microbial species shared by all individuals, stable over time (i.e., present at each time point) to identify likely candidate species which may serve as forensically-relevant targets. The core skin microbial taxa comprised of all shared species at a particular body site, together included 10 bacterial species (*Corynebacterium aurimucosum*, *Corynebacterium jeikeium*, *Corynebacterium pseudogenitalium*, *Corynebacterium tuberculostearicum*, *Micrococcus luteus*, *Propionibacterium acnes*, *Propionibacterium granulosum*, *Pseudomonas* sp., unclassified, *Rothia mucilaginosa*, *Staphylococcus epidermidis*), 1 fungal species (*Malassezia globosa*), and 1 bacteriophage (*Propionibacterium* phage P101A) (Figure 1). *Propionibacterium acnes* was the only species present in all samples at all body sites, ranging in average relative abundance from 35% to 89%, suggesting *P. acnes* may serve as an informative target species for forensic applications using skin microbiomes. Indeed, Oh et al. (9) previously reported that *P. acnes* strain single-nucleotide variant (SNV) profiles are stable and individual-specific, and the known *P. acnes* pangenome (i.e., the composition of all core and accessory genes present from all known strains of a given species) reaches saturation from all *P. acnes* strains sampled across individuals (i.e., all genes from the *P. acnes* pangenome are present across all samples). Therefore in this study, the findings of Oh et al. (9) are expanded upon and different features from *P. acnes* were evaluated as potential forensic targets in a supervised learning context.

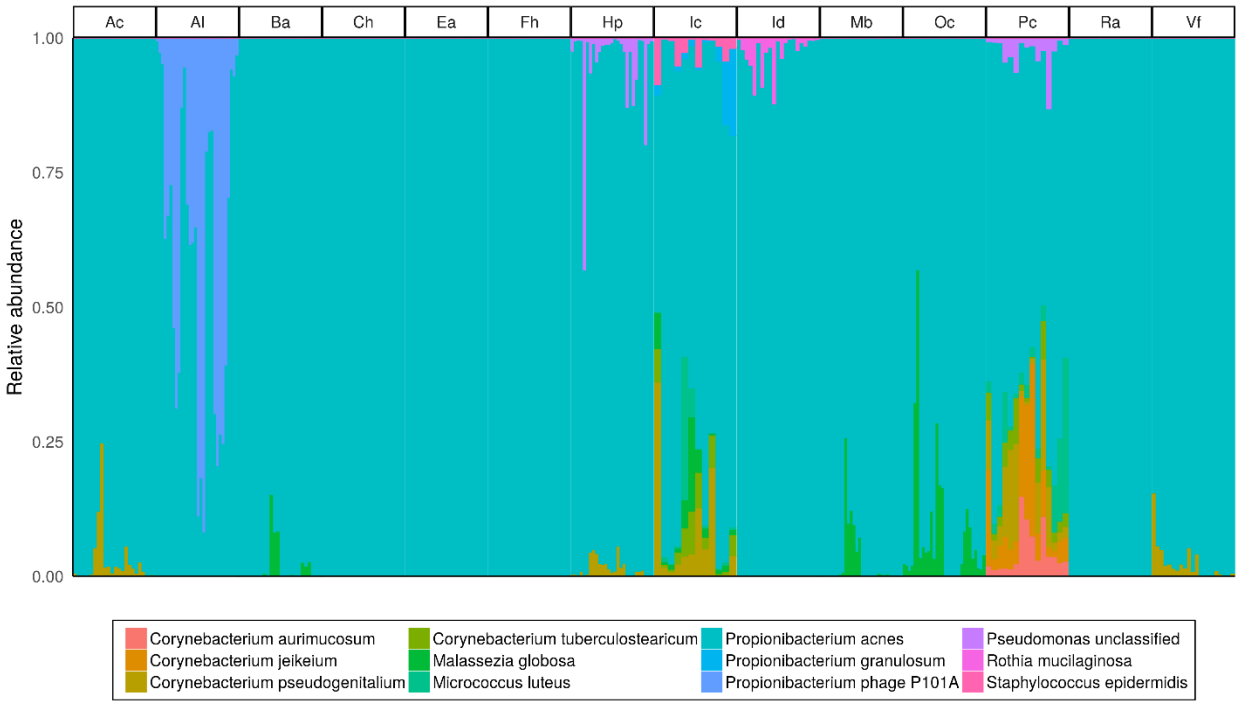

Figure 1. The proportional relative abundance of core skin microbiome taxa from 14 skin body sites. The core skin microbial taxa include prokaryotic, eukaryotic, and viral microbial species common to all samples (i.e., all individuals and time points) with ≥ 1% average relative abundance at each body site.

Propionibacterium acnes strain characterization and classification using P. acnes pangenome presence/absence features

To further assess if *P. acnes* may serve as a viable taxon for human identification, maximum likelihood phylogenetic trees were constructed over 200 markers specific to the *P. acnes* pangenome using RAxML (29). Phylogenies of *P. acnes* clade-specific markers from each individual show that *P. acnes* strains tended to place samples from the same individuals at different time points within similar positions in the tree, though some exceptions are noted (Figure 2).

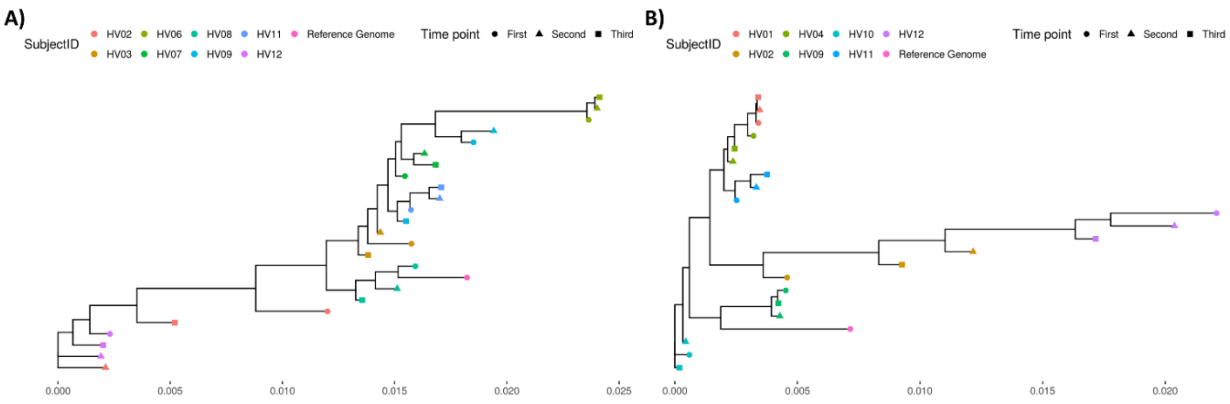

Figure 2. Maximum likelihood phylogenies of *Propionibacterium acnes* strains from all individuals and time points sampled from the A) antecubital fossa (Ac) and B) cheek (Ch). Phylogenetic trees were constructed using 200 *P. acnes* species-specific markers.

As previously reported, *P. acnes* strains across all samples reach pangenome gene saturation (9). Therefore, supervised learning using *P. acnes* pangenome gene presence/absence profiles was evaluated as a potential method for attributing skin microbiomes to their respective donors. *P. acnes* pangenome presence/absence profiles were constructed by aligning all *P. acnes* associated reads to a database comprised of all known genes from 60 *P. acnes* genomes to determine the presence or absence of each gene within each sample. Presence/absence feature vectors, comprised of 551 (ear, Ea) to 1646 (manubrium, Mb) features, were used to perform classification of host individuals across time points. In particular, regularized multinomial logistic regression (RMLR) and 1-nearest neighbor (1NN) classification (see Methods) were used to predict host individuals based on their microbiome signature taken at various time points. RMLR accuracies ranged from 66.67% at the ear (Ea) and interdigital web (Id) to 95.24% at the volar forearm (Vf) (4.67- to 9.52-fold higher accuracy than by random chance, respectively) with a mean accuracy of 79.40% (Table S2). 1NN accuracies ranged from 58.33% at the inguinal crease (Ic) to 96.30% at the hypothenar palm (Hp) (3.21- to 12.52-fold higher accuracy than by random chance, respectively) with a mean accuracy of 80.71%. RMLR and 1NN classification also were evaluated

on a reduced set of attribute selected markers (n=9 to 39), with this subset of markers chosen to have similar predictive power as the sets from which they came (see Methods). The attribute-selected loci had nearly identical classification accuracies as classification using all markers collectively (Figure 3).

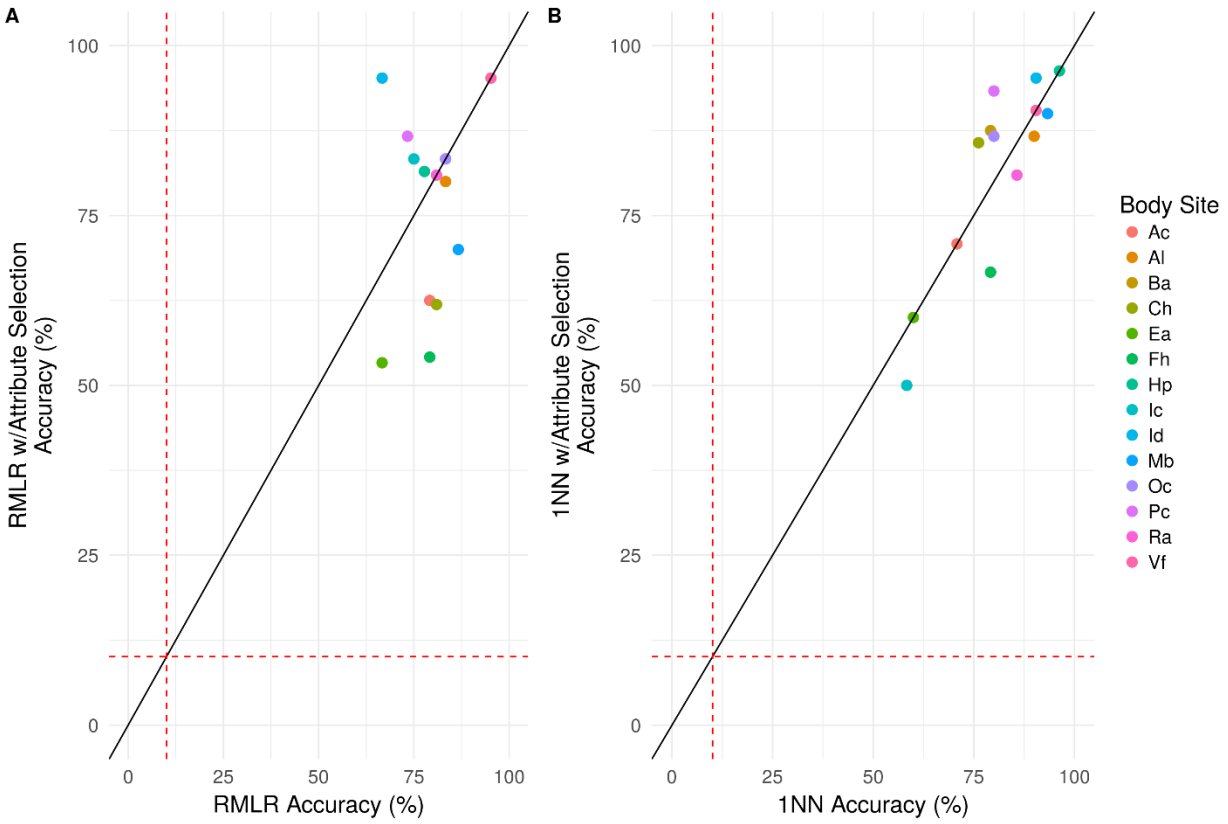

Figure 3. Classification accuracies of host individuals using *Propionibacterium acnes* pangenome gene presence/absence features. A) Regularized multinomial logistic regression (RMLR) and B) 1-nearest-neighbor (1NN) classification, with (y-axis) and without attribute selection (x-axis) were used to attribute microbiomes from three time points (spanning > 2.5 years) to their individual donor for 14 skin body sites. Red dashed lines represent the average predictive accuracy by random chance (10.1%).

Feature selection and classification of skin microbiomes using nucleotide diversities of stable clade-specific markers

The nucleotide diversities of universal, stable clade-specific markers were evaluated as a novel feature for microbiome profiling of skin microbiomes for forensic applications. Nucleotide diversity was calculated for each clade-specific marker shared by all individuals and all time points for each body site. The number of clade-specific markers shared by all samples at each body site ranged from 239 (manubrium, Mb) to 344 (popliteal fossa, Pc) markers. Principal component analysis (PCA) depicts less variation, of nucleotide diversities of all shared markers, between samples from the same individuals sampled at different times, than microbiomes from different individuals (Figure 4). As represented in Figure 4, greater variation (up to 20.85 percentage points more for the cheek, Ch) was explained by the PCA using all shared features, however marker reduction using feature selection (i.e., correlation-based feature subset selection, using the CfsSubsetEval evaluator in Weka (30); see Methods), resolves overlapping clusters from different individuals to produce more defined boundaries around samples from the same individual, likely due to the reduction of redundant features contributing towards the same level of variation.

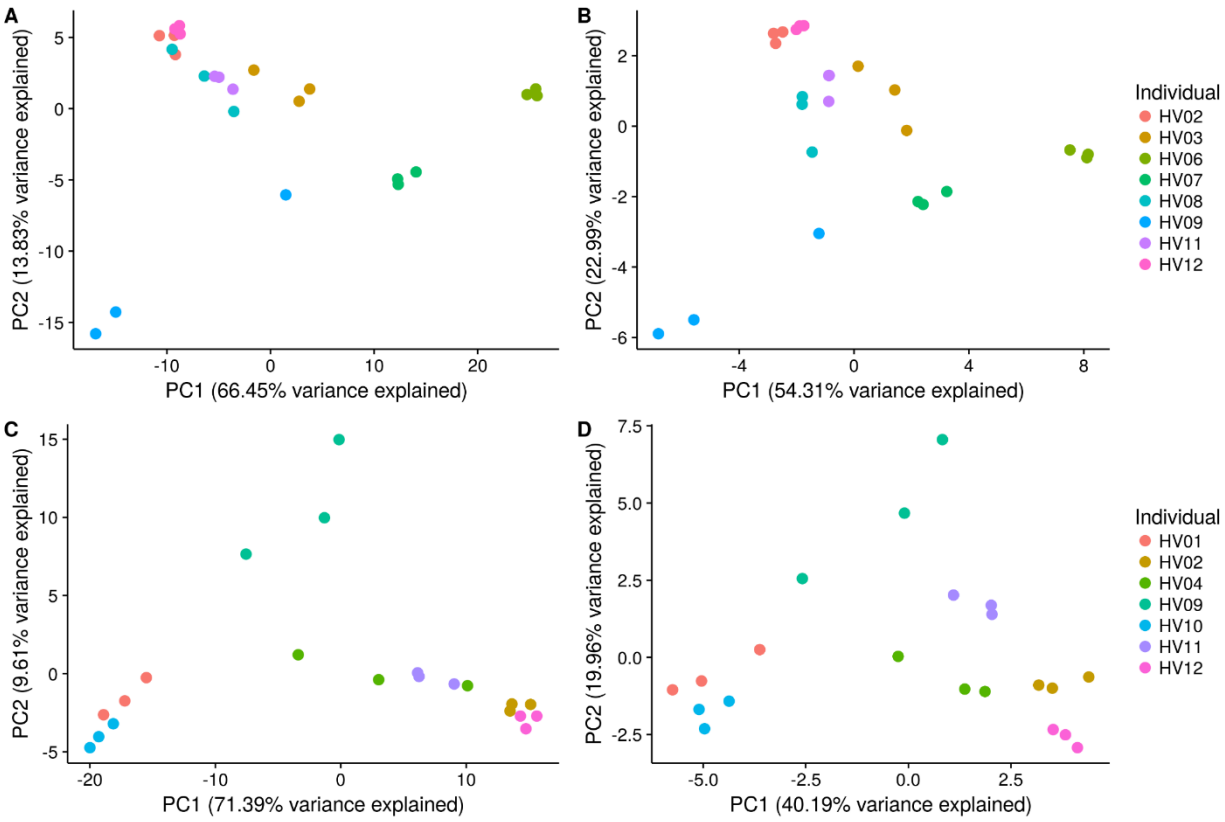

Figure 4. Principal component analysis (PCA) depicting the variance across skin microbiomes sampled from the A) antecubital fossa (Ac) and C) cheek (Ch) using the nucleotide diversity of shared clade-specific markers (242 and 252 markers, respectively) at each body site and using the nucleotide diversity from selected features using correlated feature selection to reduce the number of features to B) 27 markers at the Ac site and D) 31 markers at the Ch site.

RMLR and 1NN classification were used to classify microbiome samples with respect to their individual donor in the same manner as the assessment of presence/absence markers. RMLR accuracies ranged from 66.67% at the inguinal crease (Ic) to 100% at the cheek (Ch) (3.67- to 10-fold higher accuracy than by random chance, respectively) with a mean accuracy of 87.21% (Table S3). 1NN accuracies ranged from 56.67% at the alar crease (Al) to 100% at the inguinal crease (Ic) and popliteal fossa (Pc) (8.22- to 7-fold higher accuracy than by random chance, respectively) with a mean accuracy of 82.20%. RMLR and 1NN classification also were evaluated on a reduced set of attribute selected markers (n=14 to 47), with this subset of markers chosen to have similar

predictive power as the sets from which they came. The attribute-selected loci had nearly identical classification accuracies as classification using all markers collectively (Figure 5).

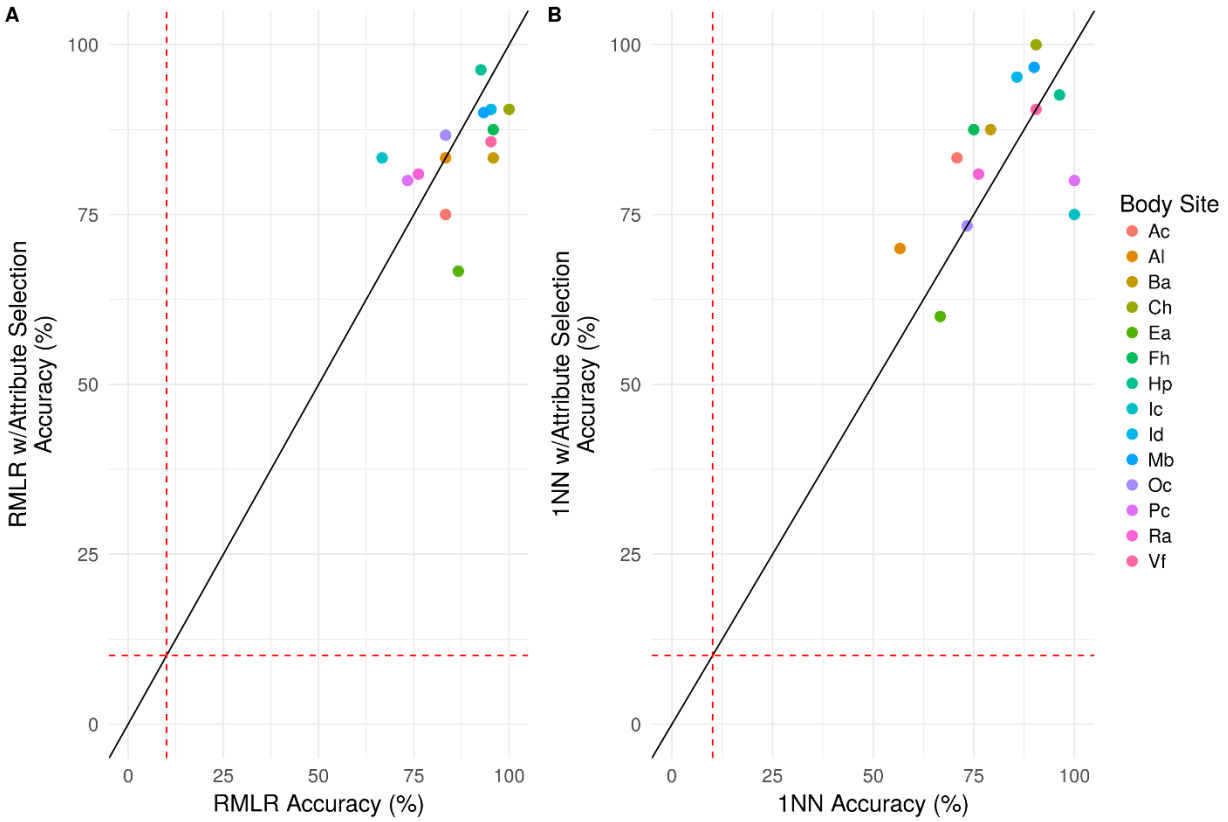

Figure 5. Classification accuracies of host individuals using nucleotide diversities of clade-specific markers shared by all individuals at each time point. A) Regularized multinomial logistic regression (RMLR) and B) 1-nearest-neighbor (1NN) classification, with (y-axis) and without attribute selection (x-axis) were used to attribute microbiomes from three time points (spanning > 2.5 years) to their individual donor for 14 skin body sites. Red dashed lines represent the average predictive accuracy by random chance (10.1%).

To assess if our classification methods were robust to differences in time, 1NN classification accuracies, with and without attribute selection, were compared between the shortest (sampling collection time points 2 vs. 3 (5-10 weeks)) and longest (sampling collection time points 1 vs. 3 (>2.5 years)) time intervals at each body site. Microbiome samples collected 5-10 weeks apart could be attributed to their host individual with higher accuracy than microbiomes samples collected >10-30 months apart (Figure S1). Long time interval accuracies ranged from 30% at the

alar crease (Al) to 100% at the popliteal fossa (Pc) and inguinal crease (Ic) with a mean accuracy of 69.52% (8.94-fold greater accuracy than by random chance). Short time interval accuracies ranged from 50% at the ear (Ea) to 100% at the forehead (Fh), inguinal crease (Ic), popliteal fossa (Pc), and volar forearm (Vf) with a mean accuracy of 85.85% (11.03-fold greater accuracy than by random chance) (Figure S1).

Feature selection identified 187 clade-specific markers from the following 12 clades that contributed the most to individual classification across all body sites: family level (n=1) (i.e, *Propionibacteriaceae*); species level (n=10) (i.e., *Corynebacterium* sp. HFH0082, *Corynebacterium tuberculostearicum*, *Propionibacterium acnes*, *Propionibacterium humerusii*, *Propionibacterium* sp. 434 HC2, *Propionibacterium* sp. 5 U 42AFAA, *Propionibacterium* sp. HGH0353, *Propionibacterium* sp. KPL1844, *Propionibacterium* sp. KPL1854, and *Propionibacterium* sp. KPL2008); subspecies level (n=1) (i.e., *Propionibacterium namnetense* SK182B-JCVI) (Table S4). These feature selected markers only represent 3 of the 12 core skin microbiome species (see Figure 1) indicating that both high- and low-abundance taxa contribute to stable features used for individual differentiation.

Assessing classifier accuracy

Several factors may influence the probability of a correct classification (*p*) of a given classifier: accuracy varied substantially across body sites (*BS*), across feature vector type (diversity or presence/absence) (*Type*), and feature selection/classifier type (*Classifier*) may also impact *p*. Conditional binomial logistic regression was used to model $\log\left(\frac{p}{1-p}\right) \sim BS + Type + Classifier$, controlling for intra-individual variation by stratifying on the (host) individual (Methods). Several of the coefficients (log odds ratios) were statistically significant (Table S5). In particular, the odds

of an accurate classification are estimated to be 28% lower for presence/absence features than for nucleotide diversity ($p<0.01$). Mean classification accuracies (p) were also contrasted between presence/absence and diversity (Figure 6) across classifier types, and as most points are above the main diagonal (i.e., higher accuracy for diversity over presence/absence), this provides further evidence that presence/absence features are less individualizing than nucleotide diversity. RMLR and 1NN, both with and without attribute selection, did not significantly impact classification accuracy. Classification accuracies did, however, significantly vary across body sites. Compared to the occiput (Oc) body site (Methods), which had medial classification accuracy, samples collected from the volar forearm (Vf) ($p<0.05$), hypothenar palm (Hp) ($p<0.01$), manubrium (Mb) ($p<0.001$), and the check (Ch) ($p<0.05$) had significantly higher odds of being classified correctly, and samples collected from the ear (Ea) ($p<0.001$) had significantly lower odds for being classified correctly (Table S5).

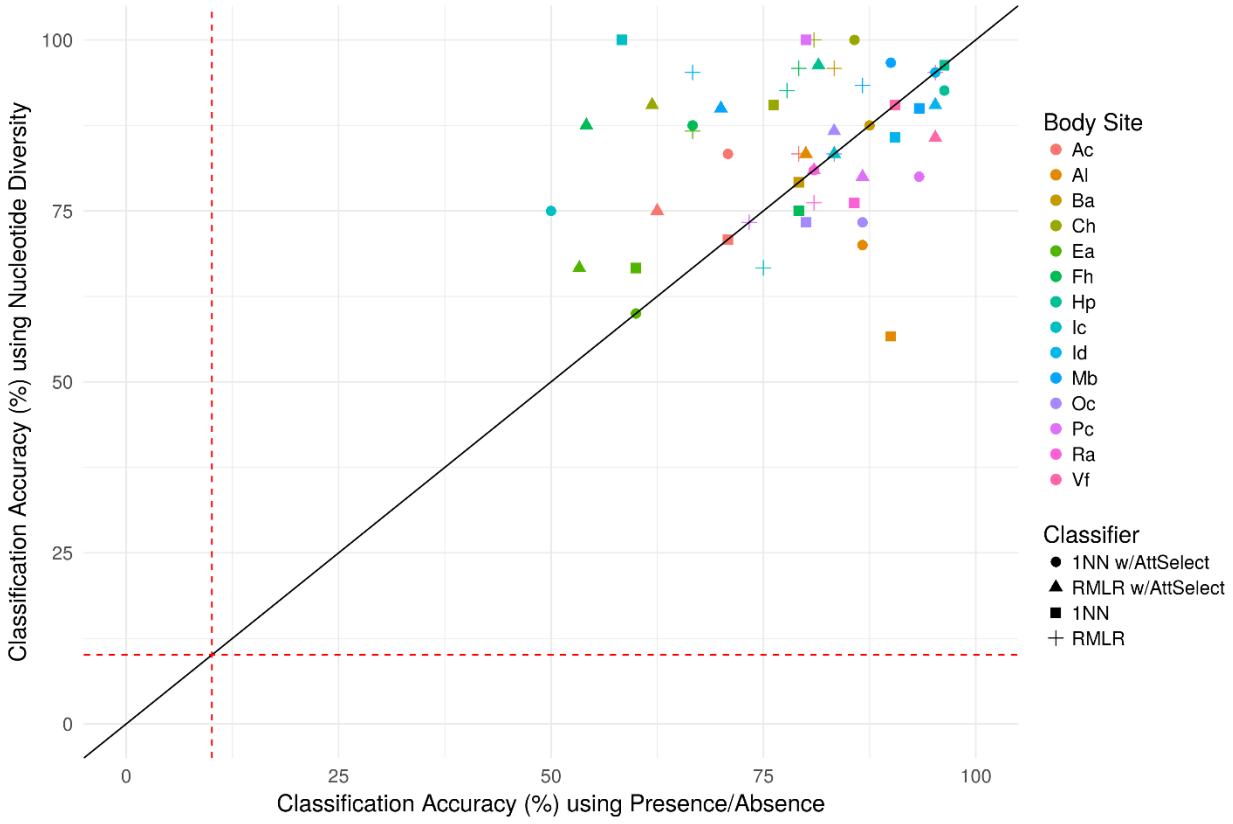

Figure 6. Comparison of classification accuracies from regularized multinomial logistic regression (RMLR) and 1-nearest-neighbor classification (1NN), with and without attribute selection (AttSelect) using *P. acnes* pangenome presence/absence features and nucleotide diversities of stable clade-specific markers. Red dashed lines represent the average predictive accuracy by random chance (10.1%).

DISCUSSION

A novel approach is described for the attribution of skin microbiome samples to their individual donors with a high degree of accuracy. Microbiome samples were collected over a large timespan (>2.5 years), and yet, classifier accuracies were high across a variety of body sites (Table S2, Table S3). Of the body sites assessed, those that are likely of the greatest forensic relevance—the Mb body site (shirt) and the Hp body site (palm)— yield highly accurate rates of classification (97%/96%, respectively, using 1NN classification on nucleotide diversity), with odds ratios of 2.64 and 2.60, respectively, relative to that of a typical body site (occiput (Oc)) (Table S5). This finding is somewhat unexpected for the hand especially as it is likely the target of frequent recolonization from life's daily tasks and has been shown to contain relatively few (~17%) shared phylotypes between different hands of the same individual (4). Lax et al. (20) observed similar classification accuracy (96.3%) when attributing microbiome samples from phone surfaces (i.e., touch samples from the hands and face) to their owners, when sampled from one time point for the majority of sample subjects and multiple time points over 2 days for 2 participants. Whereas, when assessing classification accuracy for a skin site (i.e., anterior nares) over longer time intervals (i.e., 30-300 days), Fransoza et al. (8) was only able to differentiate <30% of the total number of individuals in the study. The methods reported herein were used to attribute skin microbiomes to their hosts over long time intervals (> 2.5 years) and obtain high classification accuracy for multiple skin body sites.

In this study, two different feature types were assessed with supervised learning (i.e., RMLR and 1NN) to differentiate skin microbiomes from different individuals. *P. acnes* pangenome presence/absence features were selected based on the stability of *P. acnes* strain-level signatures and pangenome saturation over time (9) and yielded high classification accuracies (up to 96.3%), likely due to high species abundance across multiple body sites allowing for greater genome coverage for characterization. Nucleotide diversity of shared clade-specific markers was selected as a feature type to capture population-level genetic variation of stable markers, since nucleotide diversity of strains has been shown to differ significantly between individuals from different geographical regions (26) . Nucleotide diversity of stable markers yielded accuracies as high as 100% from the cheek (Ch), inguinal crease (Ic), and popliteal fossa (Pc) and contributed significantly greater (by an estimated 28%, 95% CI [10%-43%]) to classification accuracies than presence/absence features ($p<0.01$) (Table S5). This finding contrasts those from Fransoza et al. (8) which argued that minimum cardinality sets of presence/absence features (i.e., 1kb genomic window counts) are an ideal feature type for human identification. However, we demonstrate that while presence/absence features do provide high classification accuracies (Table S2), this feature type fails to capture additional genetic variation which significantly contributes to classification accuracy (i.e., nucleotide diversity) (Table S5). Furthermore, presence/absence as inferred from shotgun sequencing data are likely susceptible to stochastic effects, increasing the likelihood that informative markers may drop out in highly diverse, poorly collected, or degraded samples, sample types typical in forensic settings, and further requires parameterization on what constitutes "absence".

Attribute selection also was performed to evaluate classification performance using reduced subsets of features, selected to have similar predictive power as the full set of markers.

Since attribute selection was performed using a correlation-based approach, features were selected independent of the classifier type (unlike features selected specific to a particular classifier, e.g., (31)) and thus the markers identified in this study are potentially informative for a wide range of supervised learning algorithms. Feature selection did not have a significant effect on classification accuracy (Table S5), indicating that using an average of 24 markers reduced from 1108 for presence/absence features and an average of 32 markers reduced from 263 clade specific markers resulted in comparable classification accuracies as using full sets of features. Feature reduction helps eliminate markers which do not significantly contribute to microbiome classification (Table S5), thus eliminating potential noise and redundancy in signal, and helps select for a reduced panel of candidate markers to be developed into a multiplex assay for targeted sequencing assays for microbiome characterization.

In this study the nucleotide diversities of subsets of clade-specific markers were used to differentiate skin microbiomes samples from individuals sampled over relatively long time intervals with a high degree of accuracy. The main limitation within the study herein was sample size (n=12-30 per body site). In this study, within a given body site only three intra-individual samples were available, which limits training. Larger sample sizes are needed to further validate the methods described herein and to develop statistical models to incorporate the likelihood of microbiome classification to provide weight to similar or inclusionary comparisons. These results support future development of a robust and reproducible method for human identification using skin microbiomes. Since microbiomes likely do not have the same level of genetic stability as the human genome, identifying the most stable, personalizing features within microbiomes allows for further studies to more comprehensively assess the stability of these features and how these features contribute to classification accuracy using significantly larger population sample sets.

The study herein does not address whether the data are applicable to real or mock forensic applications (e.g., touching an object and recovering deposited skin flora). That study cannot be performed as public data of this nature are not available. More importantly performing that study would be premature. Likely for forensic applications informative targets will need to be enriched, as they are for current human identification methods. Our study has identified candidate markers that may be suitable to test forensically-relevant samples, such as touched items which would tend to have low biomass and may be somewhat degraded. Targeted enrichment and sequencing using a panel of the most informative markers would provide an ideal solution for microbiome profiling for forensic identification to obtain high coverage at stable informative sites. A multiplex is being designed to empirically test these selected candidate markers for classification accuracy and sensitivity at various sites on the human skin, including the currently low informative foot region. Once assessed for performance, larger data sets (e.g., population studies) can be generated to enable statistical weighting and resolution comparisons with those of human identification forensic genetic marker systems. The field of microbial forensics has expanded from strictly focusing on biothreat attribution to include multiple areas of microbiome applications (32), and as such, future studies should consider method development as well as new statistical models to more accurately interpret microbiome data and establish standards and validation criteria before microbiome profiling can be actively used for investigative leads and attribution within the forensic scientific community.

METHODS

Public skin microbiome dataset selection and download

Skin microbiome shotgun metagenomic datasets, comprised of samples from 12 healthy individuals across 17 body sites and sampled at different 3 time points, were downloaded using the NCBI SRA Toolkit (27), using the program fastq-dump to download 2,446 fastq files (corresponding to 585 samples) from the SRA (27), under bioproject accession PRJNA46333. Sample collection and sequencing methods are described by Oh et al. (9).

Metagenomic sequence data analysis

Metagenomic datasets were pre-processed for read quality control, using: Cutadapt (33) to remove sequence adapters, trim reads with quality scores < 20, and remove reads < 50bp; Burrows-Wheeler Alignment tool (BWA) (34) to align and remove human host-associated reads; and Samtools v1.3.1 (35) to convert sorted .bam files to fastq format for downstream use. Taxonomic classification of skin microbiomes was performed using MetaPhlAn2 (28) using default parameters. Variant calls and associated coverage for aligned MetaPhlAn2 (28) markers shared by all samples at a particular body site were determined using Samtools (35) mpileup. Only samples that met the following criteria for each body site were included in the study: ≥50x maximum coverage at any marker site within samples, ≥10x average coverage across all markers, and samples with all 3 time points for an individual (Table S1). Three body sites from the foot (i.e., plantar heel (Ph), toenail (Tn), and toe web space (Tw)) also were excluded from the study, as they only shared 2-5 markers among samples.

A custom perl script was used to parse mpileup outputs and calculate nucleotide diversity (π) of each marker, with ≥5x coverage, shared by all individuals and time points for each body site. Nucleotide diversity (π) was calculated using the following equation, $\pi = \frac{1}{n}\sum_i^n 2p_i(1-p_i)$, where p_i is the frequency of the reference base at the *i*th site in the *n*th base of the marker, as

described in Nayfach et al. (26). Strain maximum likelihood phylogenies of *Propionibacterium acnes* were constructed using RAxML (29) as implemented in StrainPhlAn (36). Briefly, StrainPhlAn was used to generate sequence alignments using MUSCLE (37), from sequence reads aligned to 200 *P. acnes* markers from MetaPhlAn2 (28), and RAxML (29) was used to generate maximum likelihood phylogenetic trees. The ggtree (38) and ggplot2 (39) R libraries using the "strainphlan_ggtree.R" script from https://bitbucket.org/biobakery/breadcrumbs was used to build the trees. Pangenome gene presence/absence profiles for *P. acnes* were generated using PanPhlAn (40), using the pre-processed "panphlan_pacnes16" database, download from https://bitbucket.org/CibioCM/panphlan/wiki/Pangenome%20databases.

Unsupervised Learning, Supervised Learning, and Attribute Selection

Principal component analysis was performed using the prcomp command in R. Statistical classification was performed in Weka (30). Classification of individuals was performed by evaluating two data feature types: nucleotide diversity and pangenome gene presence/absence. Nucleotide diversity and pangenome feature vectors were created using a custom R script, which also removed any invariant features (defined as having a standard deviation < 1e-6 across all samples). Regularized multinomial logistic regression (RMLR) and 1-nearest neighbor (1NN) classification using the Euclidean distance measure were used to perform classification, with all parameters set to their default values. Classification accuracy (i.e., the percentage of correctly classified samples in the dataset) was assessed using leave-one-out cross-validation (i.e., n-fold cross-validation; n=sample size) so as to maximize the size of the training dataset while mitigating the effects of overfitting. Thus *n* sets each composed of *n*-1 individuals were used to train classifiers, and accuracies were assessed on the single "left out" individual, with the overall

accuracies being the sums of the *n* correct and incorrect classifications. Attribute selection was performed by a correlation-based feature subset selection method, using the CfsSubsetEval evaluator in Weka (30), prior to each classification method, with default parameters and using leave-one-out cross validation. Upper and lower 95% confidence intervals were calculated for our estimates of classification accuracies using the binom.confint function from the binom R library (41) using the "asymptotic" method. All figures were created using the ggplot2 (39) and cowplot (42) R libraries unless stated otherwise. All custom scripts can be accessed at https://github.com/SESchmedes/HIDskinmicrobiome.

Conditional binomial logistic regression

Conditional binomial logistic regression was used to evaluate classifier accuracy, which models the log odds of a correct classification (p) as a linear function of the classifiers employed, the body site, and the feature vectors evaluated. In particular, $\log(\frac{p}{1-p})$ was modeled as a function of classifier type (1NN and RMLR, both with and without feature selection), the body site (column 1 of Table S1), and feature vector type, i.e., whether the classification was performed using presence/absence (encoded as 1) or diversity (encoded as 0). As these measures were repeated within individuals, traditional binomial logistic regression would otherwise underestimate error terms. Instead conditional binomial logistic regression was used to account for the repeated measures design, using the host individual as a stratum, with the clogit function in R. For the body site independent variable we chose the Oc body site as our reference category as it had medial marginal accuracy (rank 7 of 14), and the largest marginal sample size (n=240).

ACKNOWLEDGEMENTS

This project was supported by the National Institute of Justice, Award Number 2015-NE-BX-K006 and the Texas Branch of the American Society for Microbiology, 2014 Eugene and Millicent Goldschmidt Graduate Student Award. We also would like to acknowledge Jonathan L. King and David Warshauer for their support and technical assistance. We especially would like to thank the authors from Oh et al. (9) for making their skin microbiome publically available, allowing us to perform this study.

REFERENCES

1. Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260–270.

2. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

3. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. Nature 457:480–484.

4. Fierer N, Hamady M, Lauber CL, Knight R. 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci U S A 105:17994–17999.

5. Capone KA, Dowd SE, Stamatas GN, Nikolovski J. 2011. Diversity of the human skin microbiome early in life. J Invest Dermatol 131:2026–2032.

6. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, D. Lieber A, Wu F, Perez-Perez GI, Chen Y, Schweizer W, Zheng X, Contreras M, Dominguez-Bello MG, Blaser MJ. 2016. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med 8:1–14.

7. Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature 486:215–221.

8. Franzosa E a., Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower

C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci 112:E2930–E2938.

9. Oh J, Byrd AL, Park M, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854–866.

10. Hares DR. 2015. Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet 17:33–34.

11. Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 108:68–74.

12. Holland MM, Parsons TJ. 1999. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Sci Rev. 11:21-50

13. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12C:128–135.

14. Budowle B, Eisenberg AJ, van Daal A. 2009. Validity of low copy number typing and applications to forensic science. Croat Med J 50:207–217.

15. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. bioRxiv. doi: 10.1101/036103. http://biorxiv.org/content/early/2016/01/06/036103.abstract

16. Savage DC. 1977. Microbial Ecology of the Gastrointestinal Tract. Annu Rev Microbiol 31:107–133.

17. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481.

18. Goga H. 2012. Comparison of bacterial DNA profiles of footwear insoles and soles of feet for the forensic discrimination of footwear owners. Int J Legal Med 126:815–823.

19. Meadow JF, Altrichter AE, Green JL. 2014. Mobile phones carry the personal microbiome of their owners. PeerJ 2:e447.

20. Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen J a, Gilbert J a. 2015. Forensic analysis of the microbiome of phones and shoes. Microbiome 3:21.

21. Williams DW, Gibson G. 2017. Individualization of pubic hair bacterial communities and the effects of storage time and temperature. Forensic Sci Int Genet 26:12–20.

22. Nishi E, Tashiro Y, Sakai K. 2014. Discrimination among individuals using terminal restriction fragment length polymorphism profiling of bacteria derived from forensic evidence. Int J Legal Med 129:425–433.

23. Nishi E, Watanabe K, Tashiro Y, Sakai K. 2017. Terminal restriction fragment length polymorphism profiling of bacterial flora derived from single human hair shafts can discriminate individuals. Leg Med 25:75–82.

24. Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown G, Green JL, Bohannan BJ. 2015. Humans differ in their personal microbial cloud. PeerJ 3:e1258.

25. Leake SL, Pagni M, Falquet L, Taroni F, Greub G. 2016. The salivary microbiome for differentiating individuals: proof of principle. Microbes Infect 1–7.

26. Nayfach S, Pollard KS. 2015. Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria. bioRxiv DOI:10.1101/031757.

27. Kodama Y, Shumway M, Leinonen R. 2012. The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56.

28. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12:902–903.

29. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

30. Frank E, Hall MA, Witten IH. 2016. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," p. . *In* Kauffmann, M (ed.), The WEKA WorkbenchFourth Edi.

31. Tibshirani R. 1996. Regression Shrinkage and Selection via the Lasso. J R Stat Soc 58:267–288.

32. Schmedes SE, Sajantila A, Budowle B. 2016. Expansion of Microbial Forensics. J Clin Microbiol 54:1964–1974.

33. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

34. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997.

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–

2079.

36. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res 27:626–638.

37. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

38. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol Evol 8:28–36.

39. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

40. Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 13:435–438.

41. Dorai-Raj S. 2014. binom: Binomial Confidence Intervals for Several Parameterizations. R package version 1.1-1.

42. Wilke CO. 2016. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2”. R package version 0.7.0.

SUPPLEMENTAL MATERIALS

Table S1. Body sites and samples included in study

Body Site Symbol	Body Site	Body Region	Body Environment	Individuals	No. of Individuals	No. of Samples
Ac	Antecubital fossa	Arm	Moist	HV02, HV03, HV06, HV07, HV08, HV09, HV11, HV12	8	24
Al	Alar crease	Face	Sebaceous	HV01, HV02, HV03, HV04, HV05, HV08, HV09, HV10, HV11, HV12	10	30
Ba	Back	Torso	Sebaceous	HV01, HV02, HV04, HV06, HV08, HV09, HV10, HV11	8	24
Ch	Cheek	Face	Sebaceous	HV01, HV02, HV04, HV09, HV10, HV11, HV12	7	21
Ea	External auditory canal	Ear	Sebaceous	HV03, HV04, HV09, HV11, HV12	5	15
Fh	Forehead	Face	Sebaceous	HV01, HV02, HV03, HV08, HV09, HV10, HV11, HV12	8	24
Hp	Hypothenar palm	Hand	Dry	HV01, HV04, HV06, HV07, HV08, HV09, HV10, HV11, HV12	9	27
Ic	Inguinal crease	Inside Hip	Moist	HV01, HV04, HV11, HV12	4	12
Id	Interdigital web	Hand	Moist	HV01, HV04, HV08, HV09, HV10, HV11, HV12	7	21
Mb	Manubrium	Torso	Sebaceous	HV01, HV02, HV04, HV05, HV06, HV08, HV09, HV10, HV11, HV12	10	30
Oc	Occiput	Torso	Sebaceous	HV01, HV02, HV03, HV04, HV06, HV08, HV09, HV10, HV11, HV12	10	30
Pc	Popliteal fossa	Leg	Moist	HV08, HV09, HV10, HV11, HV12	5	15
Ra	Retroauricular crease	Ear	Sebaceous	HV01, HV04, HV08, HV09, HV10, HV11, HV12	7	21
Vf	Volar forearm	Arm	Dry	HV06, HV07, HV08, HV09, HV10, HV11, HV12	7	21

Table S2. Supervised learning using *P. acnes* pangenome presence/absence features

Body Site Symbol	No. of Samples	No. of Individuals	No. of Shared Markers	No. of AttSelect Markers	% Accuracy by Random Chance	RMLR				1NN				RMLR w/AttSelect				1NN w/AttSelect			
						% Accuracy	Lower 95% CI	Upper 95% CI	Ratio	% Accuracy	Lower 95% CI	Upper 95% CI	Ratio	% Accuracy	Lower 95% CI	Upper 95% CI	Ratio	% Accuracy	Lower 95% CI	Upper 95% CI	Ratio
Ac	24	8	1173	32	8.70	79.17	62.91	95.41	9.10	70.83	52.64	89.01	8.15	62.50	43.13	81.86	7.19	70.83	52.64	89.01	8.15
Al	30	10	613	27	6.90	83.33	69.99	96.66	12.08	90.00	79.26	100.70	13.05	80.00	65.68	94.31	11.60	86.67	74.50	98.83	12.57
Ba	24	8	1505	25	8.70	83.33	68.42	98.24	9.58	79.17	62.91	95.41	9.10	83.33	68.42	98.24	9.58	87.50	74.26	100.70	10.06
Ch	21	7	901	19	10.00	80.95	64.15	97.74	8.10	76.19	57.97	94.40	7.62	61.90	41.13	82.67	6.19	85.71	70.74	100.60	8.57
Ea	15	5	551	13	14.29	66.67	42.81	90.52	4.67	60.00	35.20	84.79	4.20	53.33	28.08	78.58	3.73	60.00	35.20	84.79	4.20
Fh	24	8	1047	19	8.70	79.17	62.91	95.41	9.10	79.17	62.91	95.41	9.10	54.17	34.23	74.10	6.23	66.67	47.80	85.52	7.67
Hp	27	9	1228	39	7.69	77.78	62.09	93.45	10.11	96.30	89.17	103.40	12.52	81.48	66.82	96.13	10.59	96.30	89.17	103.40	12.52
Ic	12	4	1073	9	18.18	75.00	50.50	99.49	4.13	58.33	30.43	86.22	3.21	83.33	62.24	104.40	4.58	50.00	21.71	78.28	2.75
Id	21	7	949	27	10.00	66.67	46.50	86.82	6.67	90.48	77.92	103.03	9.05	95.24	86.12	104.30	9.52	95.24	86.12	104.30	9.52
Mb	30	10	1646	24	6.90	86.67	74.50	98.83	12.57	93.33	84.40	102.20	13.53	70.00	53.60	86.39	10.15	90.00	79.26	100.70	13.05
Oc	30	10	1547	28	6.90	83.33	69.99	96.66	12.08	80.00	65.68	94.31	11.60	83.33	69.99	96.66	12.08	86.67	74.50	98.83	12.57
Pc	15	5	1142	28	14.29	73.33	50.95	95.71	5.13	80.00	59.75	100.20	5.60	86.67	69.46	103.80	6.07	93.33	80.70	105.90	6.53
Ra	21	7	981	24	10.00	80.95	64.15	97.74	8.10	85.71	70.74	100.60	8.57	80.95	64.15	97.74	8.10	80.95	64.15	97.74	8.10
Vf	21	7	1160	28	10.00	95.24	86.12	104.30	9.52	90.48	77.92	103.03	9.05	95.24	86.12	104.30	9.52	90.48	77.92	103.00	9.05

AttSelect = Attribute Selection

Table S3. Supervised learning using nucleotide diversity of shared clade-specific markers

Body Site Symbol	No. of Samples	No. of Individuals	No. of Shared Markers	No. of AttSelect Markers	% Accuracy by Random Chance	RMLR				1NN				RMLR w/AttSelect				1NN w/AttSelect			
						% Accuracy	Lower 95% CI	Upper 95% CI	Ratio	% Accuracy	Lower 95% CI	Upper 95% CI	Ratio	% Accuracy	Lower 95% CI	Upper 95% CI	Ratio	% Accuracy	Lower 95% CI	Upper 95% CI	Ratio
Ac	24	8	242	27	8.70	83.33	68.42	98.24	9.58	70.83	52.64	89.01	8.15	75.00	57.67	92.32	8.63	83.33	68.42	98.24	9.58
Al	30	10	260	42	6.90	83.33	69.99	96.66	12.08	56.67	38.93	74.39	8.22	83.33	69.99	96.66	12.08	70.00	53.60	86.39	10.15
Ba	24	8	252	30	8.70	95.83	87.83	103.80	11.02	79.17	62.91	95.41	9.10	83.33	68.42	98.24	9.58	87.50	74.26	100.70	10.06
Ch	21	7	252	31	10.00	100.00	100.00	100.00	10.00	90.48	77.92	103.00	9.05	90.48	77.92	103.00	9.05	100.00	100.00	100.00	10.00
Ea	15	5	249	26	14.29	86.67	69.46	103.80	6.07	66.67	42.81	90.52	4.67	66.67	42.81	90.52	4.67	60.00	35.20	84.79	4.20
Fh	24	8	253	37	8.70	95.83	87.83	103.80	11.02	75.00	57.67	92.32	8.63	87.50	74.26	100.70	10.06	87.50	74.26	100.70	10.06
Hp	27	9	255	47	7.69	92.59	82.71	102.40	12.04	96.30	89.17	103.40	12.52	96.30	89.17	103.40	12.52	92.59	82.71	102.40	12.04
Ic	12	4	336	22	18.18	66.67	39.99	93.33	3.67	100.00	100.00	100.00	5.50	83.33	62.24	104.40	4.58	75.00	50.50	99.49	4.13
Id	21	7	254	36	10.00	95.24	86.12	104.30	9.52	85.71	70.74	100.60	8.57	90.48	77.92	103.03	9.05	95.24	86.12	104.30	9.52
Mb	30	10	239	47	6.90	93.33	84.40	102.20	13.53	90.00	79.26	100.70	13.05	90.00	79.26	100.70	13.05	96.67	90.24	103.00	14.02
Oc	30	10	242	32	6.90	83.33	69.99	96.66	12.08	73.33	57.50	89.15	10.63	86.67	74.50	98.83	12.57	73.33	57.50	89.15	10.63
Pc	15	5	344	14	14.29	73.33	50.95	95.71	5.13	100.00	100.00	100.00	7.00	80.00	59.75	100.20	5.60	80.00	59.75	100.20	5.60
Ra	21	7	256	28	10.00	76.19	57.97	94.40	7.62	76.19	57.97	94.40	7.62	80.95	64.15	97.74	8.10	80.95	64.15	97.74	8.10
Vf	21	7	253	34	10.00	95.24	86.12	104.30	9.52	90.48	77.92	103.03	9.05	85.71	70.74	100.60	8.57	90.48	77.92	103.00	9.05

AttSelect = Attribute Selection

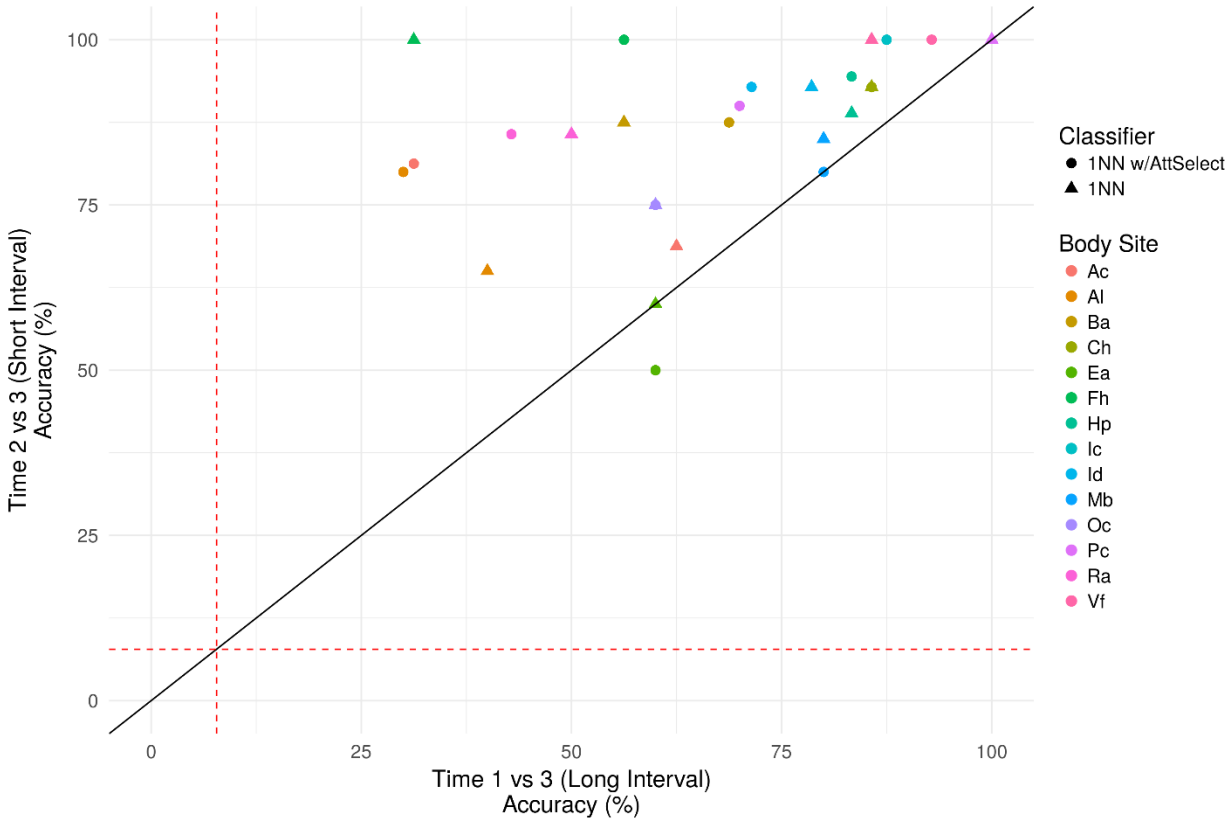

Figure S1. Comparison of 1-nearest-neightbor (1NN) classification accuracies, using leave-one-out cross validation, using the nucleotide diversities of clade-specific markers shared by all individuals from long (> 2.5 years) and short (5-10 weeks) sampling time intervals at 14 skin body sites. Red dashed lines represent the average predictive accuracy by random chance (7.8%).

Table S4. Selected features from all body sites using correlation-based feature selection

Feature	Taxonomic Clade	Level	Marker Length	Body Sites
gi|512466269|ref|NZ_KE150404.1|:c2352553-2351375	Corynebacterium_sp_HFH0082	Species	1179	Pc
gi|552867507|ref|NZ_KI515768.1|:184972-186138	Corynebacterium_tuberculostearicum	Species	1167	Pc
gi|417931402|ref|NZ_AFUN01000007.1|:c97233-97075	GCF_000221145	Subspecies	159	Vf
gi|417933187|ref|NZ_AFUN01000043.1|:4929-5147	GCF_000221145	Subspecies	219	Ra
gi|417932374|ref|NZ_AFUN01000032.1|:143771-144007	GCF_000221145	Subspecies	237	Vf
gi|335055158|ref|NZ_AFIL01000073.1|:77143-77310	Propionibacteriaceae	Family	168	Ba,Ch,Fh,Hp,Mb,Oc,Vf
gi|552896688|ref|NZ_AXMI01000003.1|:c72034-71849	Propionibacteriaceae	Family	186	Al,Mb
gi|552891898|ref|NZ_AXMG01000001.1|:c1945194-1944973	Propionibacteriaceae	Family	222	Ac,Fh,Mb,Oc
gi|552904108|ref|NZ_KI518468.1|:464070-464315	Propionibacteriaceae	Family	246	Ch,Hp,Id,Mb
gi|342211239|ref|NZ_AFUK01000001.1|:c359834-359544	Propionibacteriaceae	Family	291	Ac,Al,Mb
gi|335053539|ref|NZ_AFIL01000025.1|:23315-23623	Propionibacteriaceae	Family	309	Al,Ra,Vf
gi|335050656|ref|NZ_AFIK01000017.1|:c30516-30079	Propionibacteriaceae	Family	438	Al,Ba,Ea
gi|552879811|ref|NZ_AXME01000001.1|:c2014536-2014075	Propionibacteriaceae	Family	462	Ra
gi|355707189|ref|NZ_JH376566.1|:c170886-169537	Propionibacteriaceae	Family	1350	Ch,Ea,Fh,Ra
gi|422496709|ref|NZ_GL383802.1|:56803-56916	Propionibacterium_acnes	Species	114	Ea,Hp,Mb,Oc,Ra
gi|552876815|ref|NZ_KI515686.1|:c642879-642748	Propionibacterium_acnes	Species	132	Ea
gi|335050697|ref|NZ_AFIK01000020.1|:c12439-12299	Propionibacterium_acnes	Species	141	Ea
gi|335053685|ref|NZ_AFIL01000030.1|:c57253-57113	Propionibacterium_acnes	Species	141	Al,Pc
gi|552895565|ref|NZ_AXMI01000001.1|:c94830-94675	Propionibacterium_acnes	Species	156	Vf
gi|422539030|ref|NZ_GL384611.1|:c783227-783054	Propionibacterium_acnes	Species	174	Ra
gi|552875787|ref|NZ_KI515684.1|:c325537-325361	Propionibacterium_acnes	Species	177	Ac,Fh,Oc,Vf
gi|552891898|ref|NZ_AXMG01000001.1|:99114-99290	Propionibacterium_acnes	Species	177	Ac,Al,Hp
gi|355707384|ref|NZ_JH376567.1|:c400475-400284	Propionibacterium_acnes	Species	192	Hp,Mb
gi|552875787|ref|NZ_KI515684.1|:c488989-488798	Propionibacterium_acnes	Species	192	Ba,Ic,Oc
gi|335052413|ref|NZ_AFIK01000085.1|:c27721-27527	Propionibacterium_acnes	Species	195	Id,Pc
gi|355707384|ref|NZ_JH376567.1|:592116-592328	Propionibacterium_acnes	Species	213	Al,Ba,Fh
gi|342211239|ref|NZ_AFUK01000001.1|:535213-535428	Propionibacterium_acnes	Species	216	Ba,Ea,Hp,Id
gi|342211239|ref|NZ_AFUK01000001.1|:c1376325-1376110	Propionibacterium_acnes	Species	216	Mb
gi|342211239|ref|NZ_AFUK01000001.1|:2069064-2069282	Propionibacterium_acnes	Species	219	Ic
gi|552876418|ref|NZ_KI515685.1|:c157510-157292	Propionibacterium_acnes	Species	219	Ac
gi|422500804|ref|NZ_GL383759.1|:c166532-166311	Propionibacterium_acnes	Species	222	Ac,Ba,Fh,Hp,Ra,Vf
gi|482889214|ref|NC_021085.1|:654926-655153	Propionibacterium_acnes	Species	228	Ic,Mb
gi|355707189|ref|NZ_JH376566.1|:326756-326986	Propionibacterium_acnes	Species	231	Ac,Fh,Pc,Vf
gi|335054657|ref|NZ_AFIL01000058.1|:28786-29034	Propionibacterium_acnes	Species	249	Ac,Al,Fh,Id,Oc,Ra
gi|552876418|ref|NZ_KI515685.1|:133418-133666	Propionibacterium_acnes	Species	249	Fh,Hp
gi|552879811|ref|NZ_AXME01000001.1|:1128888-1129136	Propionibacterium_acnes	Species	249	Ac,Al
gi|552879811|ref|NZ_AXME01000001.1|:c1599141-1598893	Propionibacterium_acnes	Species	249	Mb
gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	Propionibacterium_acnes	Species	252	Mb
gi|552891898|ref|NZ_AXMG01000001.1|:1877095-1877379	Propionibacterium_acnes	Species	285	Al,Hp

gi|422434141|ref|NZ_GL384222.1|:86635-86934	Propionibacterium_acnes	Species	300	Al,Id
gi|552879811|ref|NZ_AXME01000001.1|:702826-703131	Propionibacterium_acnes	Species	306	Ac,Mb
gi|342211239|ref|NZ_AFUK01000001.1|:2001142-2001459	Propionibacterium_acnes	Species	318	Vf
gi|422482616|ref|NZ_GL383714.1|:170052-170369	Propionibacterium_acnes	Species	318	Fh,Mb,Ra
gi|335051382|ref|NZ_AFIK01000053.1|:c47134-46805	Propionibacterium_acnes	Species	330	Oc
gi|335051798|ref|NZ_AFIK01000065.1|:c4330-4001	Propionibacterium_acnes	Species	330	Ac
gi|342211239|ref|NZ_AFUK01000001.1|:c1845075-1844710	Propionibacterium_acnes	Species	366	Ch,Hp,Mb
gi|355707384|ref|NZ_JH376567.1|:621102-621467	Propionibacterium_acnes	Species	366	Al,Hp
gi|552875787|ref|NZ_KI515684.1|:c584270-583890	Propionibacterium_acnes	Species	381	Al,Oc
gi|335050749|ref|NZ_AFIK01000022.1|:c35390-34998	Propionibacterium_acnes	Species	393	Ea
gi|552891898|ref|NZ_AXMG01000001.1|:c1460921-1460529	Propionibacterium_acnes	Species	393	Ea
gi|552891898|ref|NZ_AXMG01000001.1|:793445-793843	Propionibacterium_acnes	Species	399	Ac
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	Propionibacterium_acnes	Species	399	Oc
gi|355707384|ref|NZ_JH376567.1|:c388018-387605	Propionibacterium_acnes	Species	414	Al,Ch,Fh
gi|552879811|ref|NZ_AXME01000001.1|:49241-49654	Propionibacterium_acnes	Species	414	Ic
gi|552879811|ref|NZ_AXME01000001.1|:c550719-550297	Propionibacterium_acnes	Species	423	Al,Pc
gi|552876418|ref|NZ_KI515685.1|:c713438-713010	Propionibacterium_acnes	Species	429	Hp
gi|552876418|ref|NZ_KI515685.1|:910-1341	Propionibacterium_acnes	Species	432	Ic
gi|552891898|ref|NZ_AXMG01000001.1|:834824-835255	Propionibacterium_acnes	Species	432	Id
gi|355707384|ref|NZ_JH376567.1|:190789-191232	Propionibacterium_acnes	Species	444	Al
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	Propionibacterium_acnes	Species	447	Oc
gi|552876418|ref|NZ_KI515685.1|:656232-656693	Propionibacterium_acnes	Species	462	Mb
gi|552879811|ref|NZ_AXME01000001.1|:c1651715-1651248	Propionibacterium_acnes	Species	468	Id
gi|552895565|ref|NZ_AXMI01000001.1|:619555-620031	Propionibacterium_acnes	Species	477	Ra
gi|552891898|ref|NZ_AXMG01000001.1|:c1328090-1327596	Propionibacterium_acnes	Species	495	Oc
gi|552876418|ref|NZ_KI515685.1|:c1014617-1014117	Propionibacterium_acnes	Species	501	Mb,Ra
gi|552896371|ref|NZ_AXMI01000002.1|:319095-319601	Propionibacterium_acnes	Species	507	Ba,Hp,Ic
gi|422439172|ref|NZ_GL384485.1|:c80610-80086	Propionibacterium_acnes	Species	525	Ac,Ba
gi|552895565|ref|NZ_AXMI01000001.1|:c29469-28930	Propionibacterium_acnes	Species	540	Ea
gi|552896688|ref|NZ_AXMI01000003.1|:232201-232740	Propionibacterium_acnes	Species	540	Ba
gi|552896688|ref|NZ_AXMI01000003.1|:c38494-37955	Propionibacterium_acnes	Species	540	Ch
gi|552891898|ref|NZ_AXMG01000001.1|:592123-592665	Propionibacterium_acnes	Species	543	Fh,Ic
gi|552876418|ref|NZ_KI515685.1|:36713-37258	Propionibacterium_acnes	Species	546	Ac
gi|552879811|ref|NZ_AXME01000001.1|:c1657647-1657093	Propionibacterium_acnes	Species	555	Hp,Mb,Oc,Pc
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	Propionibacterium_acnes	Species	555	Fh
gi|342211239|ref|NZ_AFUK01000001.1|:c1715790-1715233	Propionibacterium_acnes	Species	558	Ba,Mb
gi|552879811|ref|NZ_AXME01000001.1|:c2447430-2446870	Propionibacterium_acnes	Species	561	Ic
gi|552875787|ref|NZ_KI515684.1|:c96934-96368	Propionibacterium_acnes	Species	567	Al
gi|552879811|ref|NZ_AXME01000001.1|:587256-587825	Propionibacterium_acnes	Species	570	Ac,Pc
gi|552876815|ref|NZ_KI515686.1|:613740-614315	Propionibacterium_acnes	Species	576	Ba
gi|295129529|ref|NC_014039.1|:c1439020-1438442	Propionibacterium_acnes	Species	579	Ch,Ic

gi|355708280|ref|NZ_JH376568.1|:c255689-255105	Propionibacterium_acnes	Species	585	Vf
gi|552895565|ref|NZ_AXMI01000001.1|:c325088-324501	Propionibacterium_acnes	Species	588	Ch,Vf
gi|355707189|ref|NZ_JH376566.1|:507019-507612	Propionibacterium_acnes	Species	594	Ac,Ic
gi|335053761|ref|NZ_AFIL01000031.1|:46041-46637	Propionibacterium_acnes	Species	597	Hp,Id
gi|422499020|ref|NZ_GL383811.1|:10443-11039	Propionibacterium_acnes	Species	597	Hp,Mb,Oc
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	Propionibacterium_acnes	Species	600	Al,Oc
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	Propionibacterium_acnes	Species	603	Id,Mb
gi|552896371|ref|NZ_AXMI01000002.1|:638332-638937	Propionibacterium_acnes	Species	606	Id,Oc
gi|335050542|ref|NZ_AFIK01000013.1|:c12739-12119	Propionibacterium_acnes	Species	621	Ic
gi|552879811|ref|NZ_AXME01000001.1|:1327950-1328573	Propionibacterium_acnes	Species	624	Hp,Oc,Vf
gi|355708280|ref|NZ_JH376568.1|:c185858-185226	Propionibacterium_acnes	Species	633	Ba,Ea,Vf
gi|422388755|ref|NZ_GL878472.1|:c178957-178325	Propionibacterium_acnes	Species	633	Ch
gi|552895565|ref|NZ_AXMI01000001.1|:c282323-281691	Propionibacterium_acnes	Species	633	Al,Mb,Ra
gi|552895565|ref|NZ_AXMI01000001.1|:c306684-306040	Propionibacterium_acnes	Species	645	Id,Pc
gi|422386402|ref|NZ_GL878455.1|:c812899-812252	Propionibacterium_acnes	Species	648	Ac,Ch,Fh,Hp,Id
gi|552876418|ref|NZ_KI515685.1|:187493-188140	Propionibacterium_acnes	Species	648	Ch
gi|417929021|ref|NZ_AFUM01000003.1|:557611-558279	Propionibacterium_acnes	Species	669	Ea
gi|552876815|ref|NZ_KI515686.1|:c50594-49899	Propionibacterium_acnes	Species	696	Ac,Hp,Id,Vf
gi|355707189|ref|NZ_JH376566.1|:882552-883256	Propionibacterium_acnes	Species	705	Mb
gi|355707384|ref|NZ_JH376567.1|:251291-251998	Propionibacterium_acnes	Species	708	Ch,Id,Mb
gi|342211239|ref|NZ_AFUK01000001.1|:c1579497-1578787	Propionibacterium_acnes	Species	711	Hp,Ic
gi|342211239|ref|NZ_AFUK01000001.1|:1588290-1589009	Propionibacterium_acnes	Species	720	Hp
gi|552897201|ref|NZ_AXMI01000004.1|:48085-48816	Propionibacterium_acnes	Species	732	Ch,Mb,Ra
gi|335055061|ref|NZ_AFIL01000070.1|:3643-4386	Propionibacterium_acnes	Species	744	Vf
gi|552876815|ref|NZ_KI515686.1|:c586091-585333	Propionibacterium_acnes	Species	759	Al,Ic
gi|422512600|ref|NZ_GL383846.1|:26161-26922	Propionibacterium_acnes	Species	762	Mb
gi|552891898|ref|NZ_AXMG01000001.1|:1150303-1151070	Propionibacterium_acnes	Species	768	Hp,Ic,Pc,Ra
gi|552875787|ref|NZ_KI515684.1|:459339-460115	Propionibacterium_acnes	Species	777	Al,Fh
gi|552896371|ref|NZ_AXMI01000002.1|:c247178-246402	Propionibacterium_acnes	Species	777	Mb
gi|422392301|ref|NZ_GL883048.1|:64439-65218	Propionibacterium_acnes	Species	780	Vf
gi|335052272|ref|NZ_AFIK01000082.1|:c111360-110575	Propionibacterium_acnes	Species	786	Al
gi|552876418|ref|NZ_KI515685.1|:c849089-848304	Propionibacterium_acnes	Species	786	Id
gi|342211239|ref|NZ_AFUK01000001.1|:1851240-1852028	Propionibacterium_acnes	Species	789	Ac,Id
gi|335055047|ref|NZ_AFIL01000069.1|:c9632-8838	Propionibacterium_acnes	Species	795	Ic,Id,Pc
gi|387502364|ref|NC_017535.1|:c1339878-1339075	Propionibacterium_acnes	Species	804	Hp,Id
gi|335050601|ref|NZ_AFIK01000014.1|:315-1133	Propionibacterium_acnes	Species	819	Ic
gi|552879811|ref|NZ_AXME01000001.1|:368977-369813	Propionibacterium_acnes	Species	837	Al,Hp,Ic
gi|552896371|ref|NZ_AXMI01000002.1|:721564-722400	Propionibacterium_acnes	Species	837	Ch,Ic
gi|552879811|ref|NZ_AXME01000001.1|:97330-98208	Propionibacterium_acnes	Species	879	Mb
gi|422423570|ref|NZ_GL384259.1|:c300859-299957	Propionibacterium_acnes	Species	903	Fh,Hp,Ic
gi|552879811|ref|NZ_AXME01000001.1|:40840-41742	Propionibacterium_acnes	Species	903	Ac,Ba,Oc

gi|552891898|ref|NZ_AXMG01000001.1|:c2312839-2311925	Propionibacterium_acnes	Species	915	Al,Ch,Hp,Id
gi|342211239|ref|NZ_AFUK01000001.1|:527724-528653	Propionibacterium_acnes	Species	930	Al,Hp
gi|355708440|ref|NZ_JH376569.1|:c80380-79448	Propionibacterium_acnes	Species	933	Mb
gi|422538210|ref|NZ_GL384610.1|:c285619-284684	Propionibacterium_acnes	Species	936	Hp
gi|355707189|ref|NZ_JH376566.1|:1026577-1027557	Propionibacterium_acnes	Species	981	Hp
gi|552876418|ref|NZ_KI515685.1|:432422-433465	Propionibacterium_acnes	Species	1044	Ba,Fh
gi|355707384|ref|NZ_JH376567.1|:90374-91453	Propionibacterium_acnes	Species	1080	Ac,Ba,Ch,Ea,Fh,Id,Mb,Vf
gi|552879811|ref|NZ_AXME01000001.1|:1265476-1266570	Propionibacterium_acnes	Species	1095	Ac,Ea
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	Propionibacterium_acnes	Species	1116	Pc
gi|335050281|ref|NZ_AFIK01000001.1|:c2940-1807	Propionibacterium_acnes	Species	1134	Ac,Al,Ch,Fh,Hp,Oc,Ra
gi|422552858|ref|NZ_GL383469.1|:c216727-215501	Propionibacterium_acnes	Species	1227	Fh,Hp,Mb
gi|552896371|ref|NZ_AXMI01000002.1|:c671938-670697	Propionibacterium_acnes	Species	1242	Mb,Pc
gi|552879811|ref|NZ_AXME01000001.1|:c2135959-2134715	Propionibacterium_acnes	Species	1245	Ic
gi|335051382|ref|NZ_AFIK01000053.1|:c36245-34977	Propionibacterium_acnes	Species	1269	Ch,Ea,Ra
gi|342211239|ref|NZ_AFUK01000001.1|:593413-594699	Propionibacterium_acnes	Species	1287	Al,Id
gi|552897201|ref|NZ_AXMI01000004.1|:c577292-575922	Propionibacterium_acnes	Species	1371	Al,Mb
gi|552876815|ref|NZ_KI515686.1|:c200743-199319	Propionibacterium_acnes	Species	1425	Ba,Fh,Hp,Id,Oc,Vf
gi|552902190|ref|NZ_AXML01000004.1|:c579659-578172	Propionibacterium_acnes	Species	1488	Ea
gi|552876418|ref|NZ_KI515685.1|:c1032381-1030873	Propionibacterium_acnes	Species	1509	Oc,Ra
gi|335051081|ref|NZ_AFIK01000036.1|:c1716-193	Propionibacterium_acnes	Species	1524	Id,Mb
gi|335053104|ref|NZ_AFIL01000010.1|:c33862-32210	Propionibacterium_acnes	Species	1653	Al,Ch,Fh,Hp,Mb,Oc,Ra,Vf
gi|395203852|ref|NZ_AFAM01000006.1|:c75652-75533	Propionibacterium_humerusii	Species	120	Hp,Vf
gi|395205346|ref|NZ_AFAM01000017.1|:c304806-304684	Propionibacterium_humerusii	Species	123	Ba,Ea,Hp,Ra,Vf
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	Propionibacterium_humerusii	Species	126	Ba,Ch,Ea,Fh,Hp,Mb
gi|395205346|ref|NZ_AFAM01000017.1|:477016-477147	Propionibacterium_humerusii	Species	132	Ch,Oc
gi|395206455|ref|NZ_AFAM01000020.1|:c4555-4424	Propionibacterium_humerusii	Species	132	Al,Fh
gi|395206111|ref|NZ_AFAM01000018.1|:226375-226509	Propionibacterium_humerusii	Species	135	Vf
gi|395204147|ref|NZ_AFAM01000008.1|:231579-231755	Propionibacterium_humerusii	Species	177	Al,Ba,Hp,Vf
gi|395205131|ref|NZ_AFAM01000014.1|:c69464-69276	Propionibacterium_humerusii	Species	189	Fh,Oc
gi|395203469|ref|NZ_AFAM01000002.1|:37393-37605	Propionibacterium_humerusii	Species	213	Ac,Fh,Ra,Vf
gi|395203690|ref|NZ_AFAM01000005.1|:c111259-111038	Propionibacterium_humerusii	Species	222	Ba,Fh,Mb
gi|395204147|ref|NZ_AFAM01000008.1|:c721415-721191	Propionibacterium_humerusii	Species	225	Ch
gi|395205346|ref|NZ_AFAM01000017.1|:12091-12363	Propionibacterium_humerusii	Species	273	Al,Ch,Ic,Id,Vf
gi|395203061|ref|NZ_AFAM01000001.1|:c312862-312554	Propionibacterium_humerusii	Species	309	Hp,Id,Oc
gi|395203852|ref|NZ_AFAM01000006.1|:75953-76378	Propionibacterium_humerusii	Species	426	Ba,Fh
gi|395205346|ref|NZ_AFAM01000017.1|:c476952-476512	Propionibacterium_humerusii	Species	441	Id
gi|395203852|ref|NZ_AFAM01000006.1|:c137365-136916	Propionibacterium_humerusii	Species	450	Al,Ba,Ch,Ea,Fh,Hp,Mb,Oc,Ra
gi|395205346|ref|NZ_AFAM01000017.1|:c43269-42787	Propionibacterium_humerusii	Species	483	Ea,Ic
gi|395203852|ref|NZ_AFAM01000006.1|:193159-193779	Propionibacterium_humerusii	Species	621	Al,Ba,Ch,Hp,Mb,Vf
gi|395203061|ref|NZ_AFAM01000001.1|:c260639-259980	Propionibacterium_humerusii	Species	660	Ch,Ea
gi|395205131|ref|NZ_AFAM01000014.1|:c59116-58358	Propionibacterium_humerusii	Species	759	Mb,Oc

gi|395203061|ref|NZ_AFAM01000001.1|:c244616-243831	Propionibacterium_humerusii	Species	786	Ea,Vf
gi|395205346|ref|NZ_AFAM01000017.1|:c655204-654380	Propionibacterium_humerusii	Species	825	Vf
gi|395203061|ref|NZ_AFAM01000001.1|:c34216-33161	Propionibacterium_humerusii	Species	1056	Fh,Hp
gi|395203690|ref|NZ_AFAM01000005.1|:7982-10204	Propionibacterium_humerusii	Species	2223	Ac,Al,Ba,Fh,Ra
gi|335054520|ref|NZ_AFIL01000051.1|:c25042-24929	Propionibacterium_sp_434_HC2	Species	114	Al,Ba,Ch,Fh,Hp,Mb,Vf
gi|335052938|ref|NZ_AFIL01000004.1|:4461-4578	Propionibacterium_sp_434_HC2	Species	118	Oc
gi|335054139|ref|NZ_AFIL01000041.1|:c77880-77749	Propionibacterium_sp_434_HC2	Species	132	Al,Fh,Oc
gi|335053207|ref|NZ_AFIL01000016.1|:c75436-75296	Propionibacterium_sp_434_HC2	Species	141	Al,Ba,Ch,Ea,Hp,Id,Oc,Ra,Vf
gi|335054309|ref|NZ_AFIL01000044.1|:65842-65994	Propionibacterium_sp_434_HC2	Species	153	Ac,Ea,Hp,Id,Mb,Ra
gi|335055158|ref|NZ_AFIL01000073.1|:155425-155610	Propionibacterium_sp_434_HC2	Species	186	Ba,Fh
gi|335053104|ref|NZ_AFIL01000010.1|:c43071-42837	Propionibacterium_sp_434_HC2	Species	235	Ch,Ea,Id,Mb,Oc,Vf
gi|335054434|ref|NZ_AFIL01000047.1|:12103-12642	Propionibacterium_sp_434_HC2	Species	540	Fh,Mb,Oc
gi|355707189|ref|NZ_JH376566.1|:236054-236590	Propionibacterium_sp_5_U_42AFAA	Species	537	Ba,Ch,Ea,Mb,Ra
gi|514979630|ref|NZ_KE340299.1|:c1519736-1517826	Propionibacterium_sp_HGH0353	Species	1911	Id
gi|550737965|gb|AXMM01000001.1|:c339754-339413	Propionibacterium_sp_KPL1844	Species	342	Al,Ba,Fh,Hp,Id,Ra,Vf
gi|552897361|ref|NZ_AXMI01000006.1|:1-107	Propionibacterium_sp_KPL1854	Species	107	Id
gi|552897324|ref|NZ_AXMI01000005.1|:788-1104	Propionibacterium_sp_KPL1854	Species	317	Id
gi|552896371|ref|NZ_AXMI01000002.1|:767403-767774	Propionibacterium_sp_KPL1854	Species	372	Al,Ea,Fh,Hp,Id,Ra
gi|552879811|ref|NZ_AXME01000001.1|:c1820429-1820292	Propionibacterium_sp_KPL2008	Species	138	Ac,Hp
gi|552879811|ref|NZ_AXME01000001.1|:1431752-1431913	Propionibacterium_sp_KPL2008	Species	162	Id,Vf
gi|552879811|ref|NZ_AXME01000001.1|:865400-865597	Propionibacterium_sp_KPL2008	Species	198	Ra
gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	Propionibacterium_sp_KPL2008	Species	207	Al,Pc
gi|552879811|ref|NZ_AXME01000001.1|:c590861-590655	Propionibacterium_sp_KPL2008	Species	207	Id,Mb,Vf
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	Propionibacterium_sp_KPL2008	Species	270	Ba,Mb
gi|552879811|ref|NZ_AXME01000001.1|:c31864-31571	Propionibacterium_sp_KPL2008	Species	294	Al,Ch,Mb,Oc

Table S5. Conditional logistic regression odds ratios

Variable	Odds Ratio	Lower 0.95 CI	Upper 0.95 CI	P-value	
Ear (Ea)	0.38208	0.2204	0.6624	0.00061	***
Inguinal crease (Ic)	0.71288	0.3835	1.3252	0.28466	
Antecubital fossa (Ac)	0.67541	0.4002	1.1398	0.141588	
Forehead (Fh)	0.72778	0.437	1.212	0.222048	
Alar crease (Al)	1.20796	0.7315	1.9947	0.460331	
Retroauricular crease (Ra)	0.84233	0.4904	1.4468	0.534128	
Popliteal fossa (Pc)	0.72593	0.393	1.3409	0.306279	
Back (Ba)	1.56281	0.8981	2.7194	0.114143	
Cheek (Ch)	1.90783	1.0653	3.4166	0.029789	*
Manubrium (Mb)	2.64197	1.4895	4.6861	0.000891	***
Interdigital web (Id)	1.83754	0.9833	3.4341	0.056518	.
Hypothenar palm (Hp)	2.60042	1.3892	4.8678	0.002813	**
Volar forearm (Vf)	2.16785	1.0624	4.4236	0.033483	*
Presence/Absence	0.71797	0.5736	0.8987	0.003816	**
RMLR w/AttSelect	0.78406	0.5726	1.0736	0.129276	
1NN	0.84402	0.6148	1.1587	0.294287	
RMLR	1.0422	0.7527	1.443	0.803371	

Reference variables - Occiput (Oc), Nucleotide Diversity (ND), 1NN w/AttSelect

Significance codes - p<0.001 '***', p<0.01 '**', p<0.05 '*', p<0.1 '.'

CHAPTER 3

Targeted Sequencing of Clade-Specific Markers From Skin Microbiomes for Forensic Human Identification

Submitted to Forensic Science International: Genetics
2017

Sarah E. Schmedes
August E. Woerner
Nicole M. M. Novroski
Frank R. Wendt
Jonathan L. King
Kathryn M. Stephens
Bruce Budowle

ABSTRACT

The human skin microbiome is comprised of diverse communities of bacterial, eukaryotic, and viral taxa and contributes millions of additional genes to the repertoire of human genes, affecting human metabolism and immune response. Numerous genetic and environmental factors influence the microbiome composition and as such contribute to individual-specific microbial signatures which may be exploited for forensic applications. Previous studies have demonstrated the potential to associate skin microbial profiles collected from touched items to their individual owner, mainly using unsupervised methods from samples collected over short time intervals. Those studies utilize either targeted 16S rRNA or shotgun metagenomic sequencing to characterize skin microbiomes; however, these approaches have limited species and strain resolution and susceptibility to stochastic effects, respectively. Clade-specific markers from the skin microbiome, using supervised learning, can predict individual identity using skin microbiomes from their respective donors with high accuracy. In this study the hidSkinPlex is presented, a novel targeted sequencing method using skin microbiome markers developed for human identification. The hidSkinPlex (comprised of 286 bacterial (and phage) family-, genus-, species-, and subspecies-level markers), initially was evaluated on three bacterial control samples represented in the panel (i.e., *Propionibacterium acnes*, *Propionibacterium granulosum*, and *Rothia dentocariosa*) to assess the performance of the multiplex. The hidSkinPlex was further evaluated for prediction purposes. The hidSkinPlex markers were used to attribute skin microbiomes collected from eight individuals from three body sites (i.e., foot (Fb), hand (Hp) and manubrium (Mb)) to their host donor. Supervised learning, specifically regularized multinomial logistic regression and 1-nearest-neighbor classification were used to classify skin microbiomes to their hosts with up to 92% (Fb), 96% (Mb), and 100% (Hp) accuracy. All samples (n = 72) regardless of body site origin were

correctly classified with up to 94% accuracy, and body site origin could be predicted with up to 86% accuracy. Finally, human short tandem repeat and single-nucleotide polymorphism profiles were generated from skin swab extracts from a single subject to highlight the potential to use microbiome profiling in conjunction with low-biomass samples. The hidSkinPlex is a novel targeted enrichment approach to profile skin microbiomes for human forensic identification purposes and provides a method to further characterize the utility of skin microflora for human identification in future studies, such as the stability and diversity of the personal skin microbiome.

KEYWORDS: Skin microbiome · Human identification · Forensic profiling · Targeted sequencing · Supervised learning

INTRODUCTION

Diverse microbial communities of bacterial, fungal, and viral species comprise the human skin microbiome [1–3]. The skin microbiome can be influenced by several genetic and environmental factors, such as geography, health/disease states, and lifestyle (i.e., diet, hygiene, frequent contact with others, etc.) [4–8], affecting the composition of an individual's microflora. Although, a large number of skin flora are common to most individuals, overall skin microbial community profiles can vary substantially in abundance of specific microbial taxa and unique strain signatures [3,9]. Skin microbiome strain profiles can be stable over long periods of time (e.g., at least up to 3 years [3]) and thus make ideal candidates for genetically profiling microbiomes for forensic purposes.

Current forensic human identification methods typically rely on targeting autosomal markers (e.g., short-tandem repeats (STRs)) to create genetic profiles to compare evidentiary items with profiles generated from a reference sample from an individual(s) [10–16]. In some cases when the evidentiary sample may be degraded or contain low amounts of DNA (i.e., low-copy number (LCN) DNA), high-copy number (HCN) markers (e.g., the mitochondrial genome [17] or hypervariable regions of the mitochondrial genome [18,19]) are targeted. Other HCN markers, such as skin microbiome genetic markers may provide additional identifying genetic information which can be used independently or potentially in conjunction with partial human forensic marker profiles. Microbial cells transfer from the skin to objects just as with human cells, and these microbial cells are likely greater in number than human cells, ~10,000 bacterial cells/cm^2 collected per skin swab [20]. The higher number of skin microbial cells than human cells and presence of individual-specific skin microbiome signatures may make skin microbiome profiling a viable approach for potential forensic applications. However, before skin microbiome profiling can be

used for forensic human identification, a robust and reproducible method targeting stable, microbial polymorphic genetic markers must be established.

Previous studies have demonstrated the potential to use skin microbiome profiling for forensic applications, mainly targeting the 16S rRNA gene and using unsupervised methods to demonstrate that skin microbiome profiles from touched objects resemble their individual donors [21–23]. Supervised learning (i.e., classification) has been used in a limited capacity to classify skin microbiome samples from individuals collected at a single time point or over short time intervals [24,25]. Most studies characterizing skin microbiomes have relied on either targeted 16S rRNA sequencing or shotgun metagenomic sequencing; however, neither of these methods are ideal for forensic characterization of skin microbiomes due to limited species and strain resolution and susceptibility of stochastic effects, respectively. An alternative approach would be to use targeted sequencing of select sets of informative markers shown to provide individualizing resolution that are stable over time. A reliable method with the capability of strain-level resolution could be developed for forensic analyses and allow for sufficient coverage of informative sites, even from body sites with low-abundant taxa.

In a previous study, Schmedes et al. [26] mined a publically available dataset [3] comprised of shotgun metagenomic skin microbiomes collected from 12 individuals, 17 skin body sites, sampled at three time points over a time period of > 2.5 years to identify stable clade-specific markers. Markers were identified that provided individualizing resolution at each body site based on skin microbiome profiles generated using the nucleotide diversity (i.e., a measure of strain-level heterogeneity of the microbial population (See Methods)) of each marker. Supervised learning, specifically regularized multinomial logistic regression (RMLR) and 1-nearest-neighbor classification (1NN), was used to attribute skin microbiome profiles to their individual host with

high accuracy [26]. Subsets of clade-specific markers also were selected, which provide comparable classification accuracies to that of using all markers evaluated, as candidates to develop a targeted panel for skin microbiome characterization for human identification purposes [26]. Candidate markers were selected from 14/17 body sites, excluding three sites from the feet, which lacked sufficient coverage and stability for classification [26].

In this study, a novel targeted sequencing panel, the hidSkinPlex, was developed based on candidate markers from Schmedes et al. [26] for skin microbiome profiling for forensic human identification. The markers within the hidSkinPlex panel are contained in one multiplex amplification assay for targeted sequencing on the Illumina MiSeq system. Initially, the performance (i.e., sensitivity and specificity) of the hidSkinPlex was assessed using control bacterial genomic DNA from three bacterial species, *Propionibacterium acnes*, *Propionibacterium granulosum*, and *Rothia dentocariosa*. The hidSkinPlex was further evaluated using skin microbiome samples collected from three skin sites, the toe web/ball of the foot (Fb), the palm of the non-dominant hand (Hp) and the manubrium (Mb), in eight individuals. RMLR and 1NN classification were used to predict skin microbiomes originating from specific body sites with their respective donors. Attribute selection also was performed to identify subsets of hidSkinPlex markers that provide similar or greater predictive power than the entire hidSkinPlex panel for individual classification at each body site. Additionally, maximum likelihood phylogenies of *P. acnes* strains, using *P. acnes*-specific markers from the hidSkinPlex were constructed to characterize *P. acnes* strains across body sites and individuals to determine if *P. acnes* strains were more related at the level of the individual or the individual at each body site. Finally, hidSkinPlex profiles and human-specific STR and single-nucleotide polymorphism (SNP) profiles generated from the same skin samples were compared to provide a case study on the

potential to use skin microbiome profiles in conjunction with human genetic profiles for forensic investigative purposes.

MATERIALS AND METHODS

Sample collection

Skin microbiome samples were collected from eight individuals (four females, four males) sampled from the Mb, Hp, and Fb, according to a protocol approved by the University of North Texas Health Science Center (UNTHSC) Internal Review Board (IRB). Skin microbiome samples were collected using 4N6FLOQSwabs™: Genetics (COPAN, Brescia, Italy) pre-moistened with 30 µL sterile, molecular-grade water (Phenix, Candler, NC). All skin swabs were collected by swabbing a separate section of skin per replicate with firm pressure for 10 seconds on one side of the swab head, rotated 180°, and then swabbed another 10 seconds. Mb skin sites were collected ~5 cm beneath the junction of the clavicles. Hp samples were collected by swabbing separate sections of the palm starting at the base of a finger (excluding the thumb) and extending across the entire length of the palm. Fb samples were collected by swabbing between each toe web space and extending down the entire length of the ball of the foot. Three replicate samples were collected from each body site for a total of nine swabs collected per individual (n = 72). Each subject filled out a questionnaire to retrieve associated metadata related to the subject regarding bioancestry, hygiene, health/disease state, and recent travel. No subjects were eliminated from the study due to answers on the questionnaire. Swabs were either stored at -20 °C until DNA extraction or extracted directly.

DNA extraction and quantification

Total DNA was extracted from skin swabs collected from subjects S001-S004 using the MO BIO BiOstic® Bacteremia DNA Isolation Kit (MO BIO Laboratories, Inc. Carlsbad, CA) following the manufacturer's protocol with the following modifications. The CB1 buffer was added directly to the MicroBead tube followed by adding the swab to the buffer/bead solution and allowed to soak for 5 minutes with occasional rotation of the swab. Next, the swab head was snapped off, along the break point on the plastic applicator, and left in the tube proceeding to the 70 °C incubation step; the remainder of the manufacturer's protocol was followed as prescribed. A swab blank was included with each extraction. DNA extracts were stored at -20 °C. Total DNA was extracted from skin swabs collected from subjects S005-S008 using the QIAamp BiOstic Bacteremia DNA Kit (Qiagen, Hilden, Germany), the new version of the previous MO BIO kit. The same modified swab protocol was followed except the swab head could not fit in the new PowerBead tube. Instead, the swab was soaked with agitation in the MBL buffer (previously CB1) for at least 5 minutes followed by adding the supernatant from the swab tube directly to the PowerBead tube, proceeding to the 70 °C incubation step; the remainder of the manufacturer's protocol was followed as prescribed. Total DNA was quantified using the Qubit® 2.0 Fluorometer with the Qubit® dsDNA HS Assay Kit (ThermoFisher Scientific, Eugene, OR).

Development of the hidSkinPlex panel and multiplex primer design

Markers included in the hidSkinPlex panel were selected by Schmedes et al. [26]. Briefly, publically-available shotgun metagenomic sequence datasets generated by Oh et al. [3] were mined to identify universal clade-specific markers (i.e., markers unique to a particular microbial taxonomic clade) that were stable over the tested time interval, which could be used to differentiate

individuals based on their individual-specific skin microbiome signatures. The Oh et al. [3] data were comprised of skin microbiomes from 12 healthy individuals, 17 skin body sites, and 3 time points (sampled over a period of > 2.5 years). The nucleotide diversities of clade-specific markers, from the MetaPhlAn2 [27] database, common to all individuals and time points at each body site, were calculated and used as features with RMLR and 1NN classification with and without attribute selection (e.g., correlation-based feature selection) to attribute skin microbiomes to their respective host donors. Attribute selected markers (i.e., a subset of markers with comparable predictive power as all shared markers) were included in the hidSkinPlex panel. Markers included in the hidSkinPlex panel identified in Schmedes et al. [26] were selected from samples which met the following criteria for sample inclusion: ≥ 50x maximum read depth at any shared marker site, ≥ 10x average read depth for all shared markers, and detected in all 3 time points for each individual per body site. Marker sites were included for analysis using a threshold of ≥ 5x read depth. Additional attribute selected markers, which were not selected by Schmedes et al. [26], were included in the hidSkinPlex panel to build in redundancy in the panel in case particular markers failed to amplify. Additional markers were selected using marker inclusion thresholds of ≥ 2x and ≥ 10x read depth and an additional sample set (≥ 30x maximum read depth at any shared marker site, ≥5x average read depth for all shared markers, and detected in all 3 time points for each individual per body site) with marker inclusion thresholds of ≥ 2x and ≥ 5x read depth at each marker site. The final hidSkinPlex panel contained 286 markers from 22 bacterial (and phage) clades (Table S1).

Custom primers (n = 572) for each hidSkinPlex marker (n = 286) were designed by Verogen, Inc. Primers were designed to produce amplicons with maximum coverage across each marker reference sequence with no overlapping primers. Primers for amplicons less than 200 bp

also incorporated the Nextera Transposase sequence (Illumina, Inc., San Diego, CA) on the 5' end of the primer to ensure transposition during library preparation.

Development and evaluation of the hidSkinPlex multiplex assay

The hidSkinPlex amplification assay was developed using the QIAGEN® Multiplex PCR Plus Kit (Qiagen) and three bacterial DNA controls (*P. acnes* Strain SK137, *P. granulosum* D-34, and *R. dentocariosa* Strain M567) (ATCC, Manassas, VA). The quantities of total bacterial genomic DNAs were determined using the Qubit® 2.0 Fluorometer with the Qubit® dsDNA BR Assay Kit (ThermoFisher Scientific). Each of the custom primers were at 100 µM final concentration (Integrated DNA Technologies, Coralville, IA) and pooled to make a working stock, 175 nM for each primer. Multiplex reaction conditions, following recommendations in the protocol for "Multiplex PCR fragments up to 1.5 kb in length" [28], were as follows for a 50 µL reaction: 25 µL Multiplex PCR Master Mix; 5 µL 10x primer mix; 5 µL 5x Q-Solution (with and without); 1 ng template DNA; molecular-grade water (volume varies according to volume of sample added). Separate PCRs with the following modified conditions were evaluated: 17.5 nM, 8.75 nM, and 4.375 nM final primer concentrations, with and without the addition of Q-Solution, 1 ng each control DNA and a 1:1:1 mixture including each bacterial control sample (1 ng total). PCR conditions were as follows: 95 °C for 5 minutes; 40 cycles (95 °C for 30 seconds; 55 °C, 57 °C or 59 °C for 3 minutes; 72 °C for 90 seconds); and 68 °C for 10 minutes. PCR product was purified using the MinElute® PCR Purification Kit (Qiagen) and quantified using the Qubit® 2.0 Fluorometer with the Qubit® dsDNA BR Assay Kit (ThermoFisher Scientific). Purified PCR product was visualized on the 2200 TapeStation system (Agilent Technologies, Santa Clara, CA) with the D1000 ScreenTape and reagents (Agilent Technologies) or with the High Sensitivity

D1000 ScreenTape and reagents (Agilent Technologies) (using a 1:20 dilution of purified PCR product).

Library preparation and hidSkinPlex targeted sequencing

Targeted hidSkinPlex sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina) with the Nextera XT Index Kit v2 Set C (Illumina) and 1% spiked-in PhiX Control v3 (Illumina), following manufacturer's protocol, using 90 µL volume of Agencourt® AMPure® XP beads (Beckman Coulter, Inc., Brea, CA) during library cleanup. Libraries were quality controlled and visualized on the 2200 TapeStation system (Agilent Technologies) with the High Sensitivity D1000 ScreenTape and reagents. Pooled libraries were sequenced on the MiSeq (Illumina) using the MiSeq Reagent Kit v2 (300-cycles) (Illumina) with a 2x150 bp read length.

DNA extracts from S001, including a reference buccal swab, also were analyzed using the ForenSeq™ DNA Signature Prep Kit (Primer Mix A) (Illumina) and sequenced on the MiSeq FGx™ Forensic Genomics System (Illumina), following manufacturer's instructions. ForenSeq data were analyzed using STRait Razor v2s [29].

Sequence quality control and data analysis

Sequence data were preprocessed using cutadapt [30] to trim bases with a quality score less than 20 and remove reads less than 50 bases in length. Adapters were previously removed on the MiSeq system before data analysis. MetaPhlAn2 [27] was used to align sequence reads to the MetaPhlAn2 reference database, which includes the markers in the hidSkinPlex panel. Samtools [31] programs view, sort, stats, index, bedcov and mpileup were used to retrieve alignment

statistics and calculate read depth and variant calls for each aligned marker in the hidSkinPlex panel. To assess the performance of the hidSkinPlex, accuracy calls (i.e, true positive (TP), true negative (TN), false positive (FP), false negative (FN)) were designated by the following criteria: TP = expected marker, present; TN = expected absent marker, absent; FP = expected absent marker, present; FN = expected marker, absent. The sensitivity (SN = $\frac{TP}{TP+FN}$) and specificity (SP = $\frac{TN}{TN+FP}$) of the hidSkinPlex were calculated for accuracy calls using a threshold of 70x read depth (i.e., > maximum read depth observed in the reagent blank).

Custom perl and R scripts were used to calculate the nucleotide diversity (π), $\pi = \frac{1}{n}\sum_i^n 2p_i(1-p_i)$, where p_i is the frequency of the reference base at the *i*th site in the *n*th base of the marker (as described in Nayfach et al. [32]) of each marker and construct feature vectors to use for statistical classification. Classification was performed to attribute skin microbiome profiles to their individual hosts using RMLR and 1NN in Weka [33] using *n*-fold cross validation where *n* is the sample size and the training set is of size *n* - 1 (i.e., "leave-one-out cross-validation" (LOOCV)). LOOCV helps provide precise estimates of prediction accuracy by testing each sample against a maximally-sized training set, minus the test sample, while mitigating the effects of overfitting. Attribute selection, using the CfsSubsetEval in Weka [33], also was performed prior to each classification method to select for a subset of markers which may have similar weight than using the full set. Subsets of markers were evaluated since hidSkinPlex markers were selected across 14 body sites at different read depth thresholds, potentially building in marker redundancy and markers performing best at particular body sites. Therefore, subsets of markers may be better suited for classification at specific body sites. Upper and lower 95% confidence intervals on the binomial probability of the classification accuracy estimates were calculated using the bionom.confint in the binom R library [34] using the asymptotic method. Fisher's Exact tests were

performed in R using the fisher.text function. All figures were made in R using the ggplot2 [35] and cowplot [36] R libraries, unless otherwise stated.

Principal components analysis (PCA), using the nucleotide diversities of the hidSkinPlex markers, was performed using the prcomp function in R. Maximum likelihood phylogenies of hidSkinPlex *P. acnes* species-specific markers were constructed using MUSCLE [37] and RAxML [38] as implemented in StrainPhlAn [39] and the "strainphlan_ggtree.R" script from https://bitbucket.org/biobakery/breadcrumbs using the ggtree [40] and ggplot2 [35] R libraries.

Data and script accessibility

Sequence datasets can be found on the NCBI Sequence Read Archive under BioProject ID accession PRJNA398026. Custom perl and R scripts can be accessed at https://github.com/SESchmedes/hidSkinPlex.

RESULTS

Development and evaluation of the hidSkinPlex targeted sequencing assay

The hidSkinPlex panel consists of 286 clade-specific markers from 22 bacterial (and phage) clades selected from the MetaPhlAn2 [27] reference database (Table S1), with > 65% of the markers from the dominant skin flora, *P. acnes*. Primers were designed to maximize coverage of each panel marker, without tiling, producing 286 amplicons (n = 572 primers) ranging in size from 72 bp to 721 bp (average 464 bp) (Figure 1). The percentage of the marker reference sequences covered ranges from 32% to 100% (average 82%) with two amplicons designed with lengths greater than the reference genomic region. Nextera transposase sequences were incorporated into primers for amplicons < 200 bp to improve tagmentation efficiency during

library preparation. Multiplex parameters including, the annealing temperature (i.e., 55 °C, 57 °C, and 59 °C), primer concentration (i.e., 17.5 nM, 8.75 nM, and 4.375 nM, each), and use of Q solution (QIAGEN) were evaluated to test the performance of the panel on 1ng of bacterial control genomic DNA from *P. acnes* Strain SK137, *P. granulosum* D-34, and *R. dentocariosa* Strain M567 which include at least 200 markers from the hidSkinPlex panel. The hidSkinPlex also was assessed on a 1:1:1 mixture (1 ng total) of each bacterial control. Initially, the hidSkinPlex was evaluated using 17.5 nM primer concentration with an annealing temperature of 57 °C. However, primer-dimer concentrations were elevated (data not shown) and lower primer concentrations of 8.75 nM and 4.375 nM were used for a 3-stage temperature gradient (e.g., 55 °C, 57 °C, and 59 °C) assessment of the multiplex. Samples amplified using the following conditions were evaluated through the full sequencing workflow, based on amplification assessment on the Agilent 2200 TapeStation: 57 °C and 59 °C annealing temperatures; 8.75 nM primer concentration with and without Q solution (data not shown); and 4.375 nM primer concentration without the addition of Q solution (data not shown).

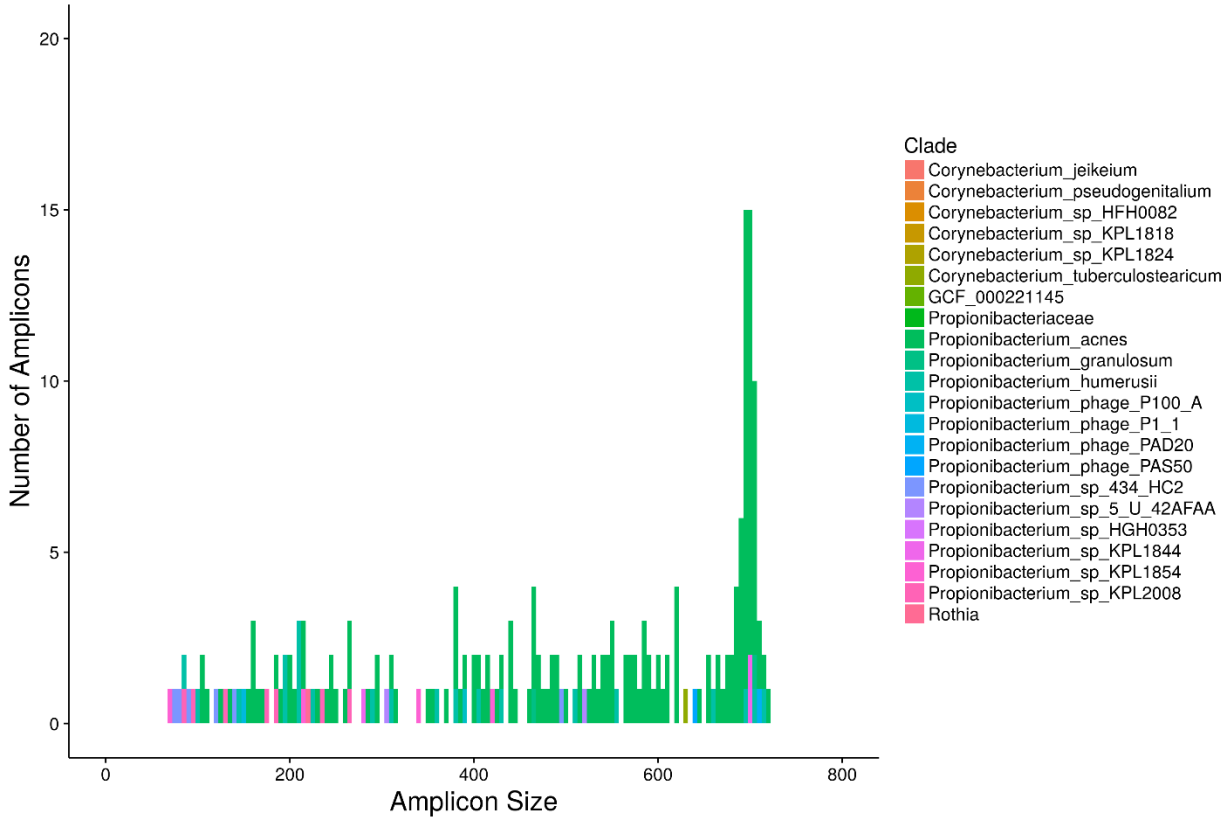

Figure 1. A histogram of the amplicon sizes present in the hidSkinPlex panel. (Bin size = 5).

A total of 22.5 million raw sequencing reads (average ~0.94 million reads per sample) were generated with 14 million reads (average ~0.58 million reads per sample) remaining after quality trimming and filtering. A total of 242 markers amplified and were detected by sequencing; however, after implementing a threshold of ≥ 70x read depth, 200/200 expected markers were detected and were sequenced with average read depths (computed by total read depths at each base across the marker/length of amplicon) per marker ranging from 70x to > 49,000x read depth with an average of 1,278x ± 2,276 (SD) read depth (all reads in the reagent blank were < 70x and likely due to low-level bacterial contaminants from reagents [41,42]) (Figure 2A). The performance of the panel was assessed by determining the proportion of true positives, false positives, true negatives, and false negatives based on expected marker presence/absence and calculating the

sensitivity and specificity of the hidSkinPlex (Figure 3A, Table S2). The proportion of true positives and true negatives ranged from < 30% to > 85% (Figure 3A); however, after implementing a threshold of ≥ 70x read depth the proportion of expected accuracy ranged from > 85% to 100% (Figure 3B). The sensitivity of the hidSkinPlex panel, for 200/286 markers at ≥ 70x read depth, ranged from 85% - 98% with a specificity range of 76% - 90% (Table S2). Average read depth across each marker for true positives ≥ 70x read depth ranged from 70x to > 33,000x read depth (average of 1,123x ± 1,508 (SD) read depth) (Figure 2B). PCR parameters of 59 °C and 8.75 nM primer concentration without Q solution (QIAGEN) were selected to assess the performance of the hidSkinPlex on skin microbiome samples since these parameters resulted in overall higher and more uniform read depth of expected markers evaluated on each control sample (Figures 2-3, S1). More weight was given to the performance of *P. acnes* (n = 196) and the synthetic bacterial mixture (n = 200), since most of the markers in the panel cover this species/sample as opposed to *P. granulosum* (n = 12) and *R. dentocariosa* (n = 1).

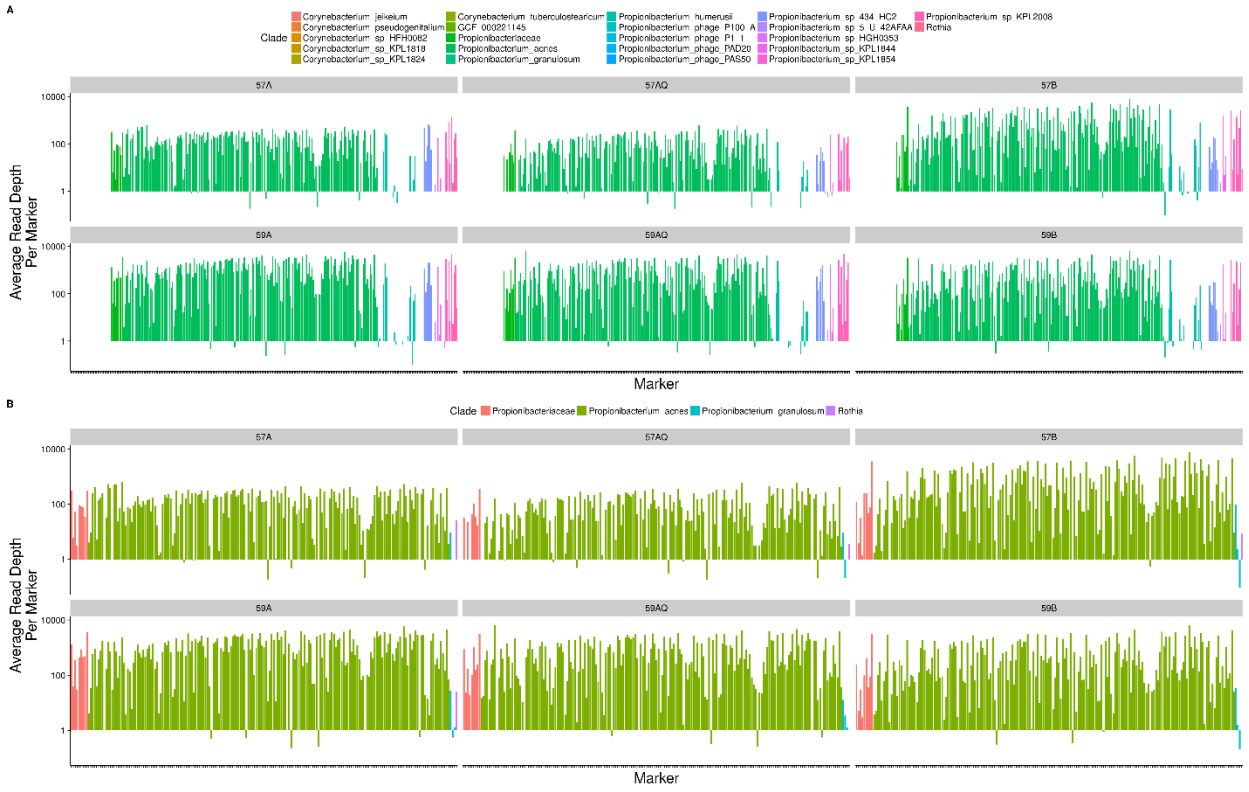

Figure 2. The average read depth at each hidSkinPlex marker. A) Marker read depth at each marker in the hidSkinPlex (n = 286) for a synthetic bacterial mixture containing equal amounts of genomic DNA from *Propionibacterium acnes*, *Propionibacterium granulosum*, and *Rothia dentocariosa*. B) Marker read depth at each expected marker (i.e., "true positive", n = 200) for a synthetic bacterial mixture containing equal amounts of genomic DNA from *P. acnes*, *P. granulosum*, and *R. dentocariosa*. PCR parameters tested, include: 57°C and 59°C annealing temperatures; A = 8.75 nM final primer concentration; B = 4.375 nM final primer concentration; Q = addition of Q solution. (Markers ordered by clade then amplicon size for each PCR multiplex parameter, on a log scale).

Figure 3. Performance of the hidSkinPlex on bacterial controls *Propionibacterium acnes*, *Propionibacterium granulosum*, and *Rothia dentocariosa*. Accuracy calls (i.e, true positive (TP), true negative (TN), false positive (FP), false negative (FN)) were designated by the following criteria: TP = expected marker, present; TN = expected absent marker, absent; FP = expected absent marker, present; FN = expected marker, absent. A) Proportion of accuracy calls using a threshold of 1x read depth. B) Proportion of accuracy calls using a threshold of 70x read depth (i.e., > maximum read depth observed in the reagent blank). Sample names: BacMix = synthetic bacterial mixture containing equal amounts of genomic DNA from *P. acnes*, *P. granulosum*, and *R. dentocariosa*; Pacnes = *P. acnes*; Pgran = *P. granulosum*; RB = reagent blank; Rdent = *R. dentocariosa*. PCR parameters tested, include: 57°C and 59°C annealing temperatures; A = 8.75 nM final primer concentration; B = 4.375 nM final primer concentration; Q = addition of Q solution.

Skin microbiome profiling and classification using the hidSkinPlex

The hidSkinPlex was evaluated on skin microbiome samples to assess if enrichment of targeted clade-specific markers can be used to differentiate individuals based on microbiome profiles. Skin microbiome samples were collected from eight individuals, sampled in triplicate

from Mb, Hp, and Fb (n = 72 samples). The Mb and Hp body sites were selected for this study due to their forensic relevance (i.e., Mb (shirt collar) and Hp (touch items)) and to overlap sites previously tested by Schmedes et al. [26], where the classification accuracies were generally higher. The foot was selected to determine if skin microbiomes from the foot can be used to differentiate individuals using targeted enrichment of informative hidSkinPlex markers. Previous attempts to use skin microbiome profiles using shotgun metagenomic data from the foot were not possible [26] due to low sequence read depth and/or coverage and high variability of markers at the foot body site [3].

DNA extracts (50 µL total volume) from the collected skin microbiome samples generated quantification results of total DNA ranging from < 0.5 to 934 pg/µL. A total of 1 ng of DNA template or up to 20 µL (maximum volume) of DNA template for each sample was amplified using the hidSkinPlex and sequenced generating 122 million raw sequencing reads (average of ~1.7 million reads per sample). Sequence reads were preprocessed to remove sequence adapters, trim bases with a quality score < 20 and remove reads < 50 bases in length resulting in 91.2 million total sequence reads (average ~1.3 million reads per sample) for downstream analysis. Sequence reads aligned to 282 out of the 286 total markers in the hidSkinPlex panel with read depths per marker ranging from 0.07x (less than 100% of the marker captured) to > 64,000x read depth (average of 2,117x ± 6,305 (SD) read depth) (Figures 4, S2-4). A total of 183 markers, termed hereafter as universal markers, were common to all individuals and all body sites with a minimum of 2x read depth.

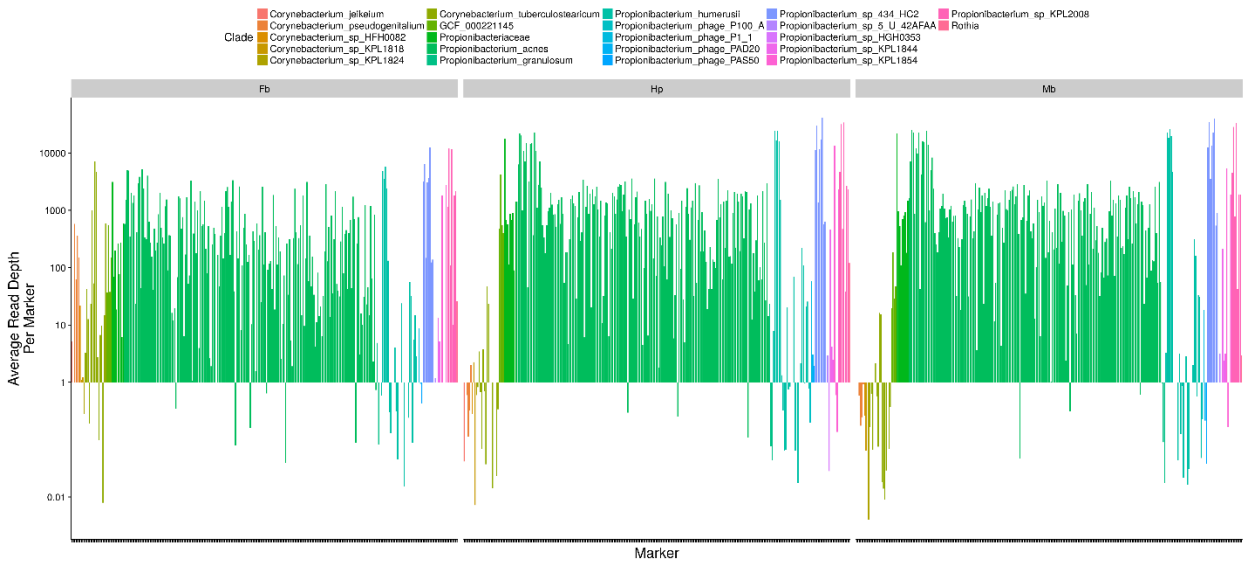

Figure 4. The average read depth at each hidSkinPlex marker present in eight individuals from the toe web/ball of the foot (Fb), palm of the non-dominant hand (Hp) and manubrium (Mb). Markers are ordered by clade then amplicon size on a log scale.

To assess the ability of select subsets of hidSkinPlex markers to differentiate skin microbiomes from different individuals, skin microbiome profiles were constructed by calculating the nucleotide diversity for each marker (See Methods). Marker nucleotide diversity captures the level of heterozygosity of each marker and can capture strain level variation [26,32]. Nucleotide diversities were calculated for seven read depth thresholds (i.e., 2x, 10x, 25x, 50x, 100x, 150x, 200x) for samples at each body site and all sites combined. Classification was performed for all body site samples combined to test the prediction accuracy when the body site is unknown to the classifier, in contrast to previous studies [9,24–26] in which the body site was known (i.e., conditioning on the body site). Skin microbiome profiles were assessed using subsets of universal (i.e., markers common to all individuals and body sites, including all replicates) and non-universal markers (i.e., all markers present across all samples, including common and unique markers). PCA of skin microbiomes profiles using universal markers depicted samples from the same individual at Fb, Hp, and Mb body sites clustering more closely than samples from different individuals, with

few exceptions (Figure 5). This cluster pattern was less apparent when considering all samples together, regardless of body site; however, some clustering was still observed (Figure 5). While unsupervised learning, such as PCA, can facilitate data visualization, supervised methods are necessary to calculate predictive accuracies for sample classification.

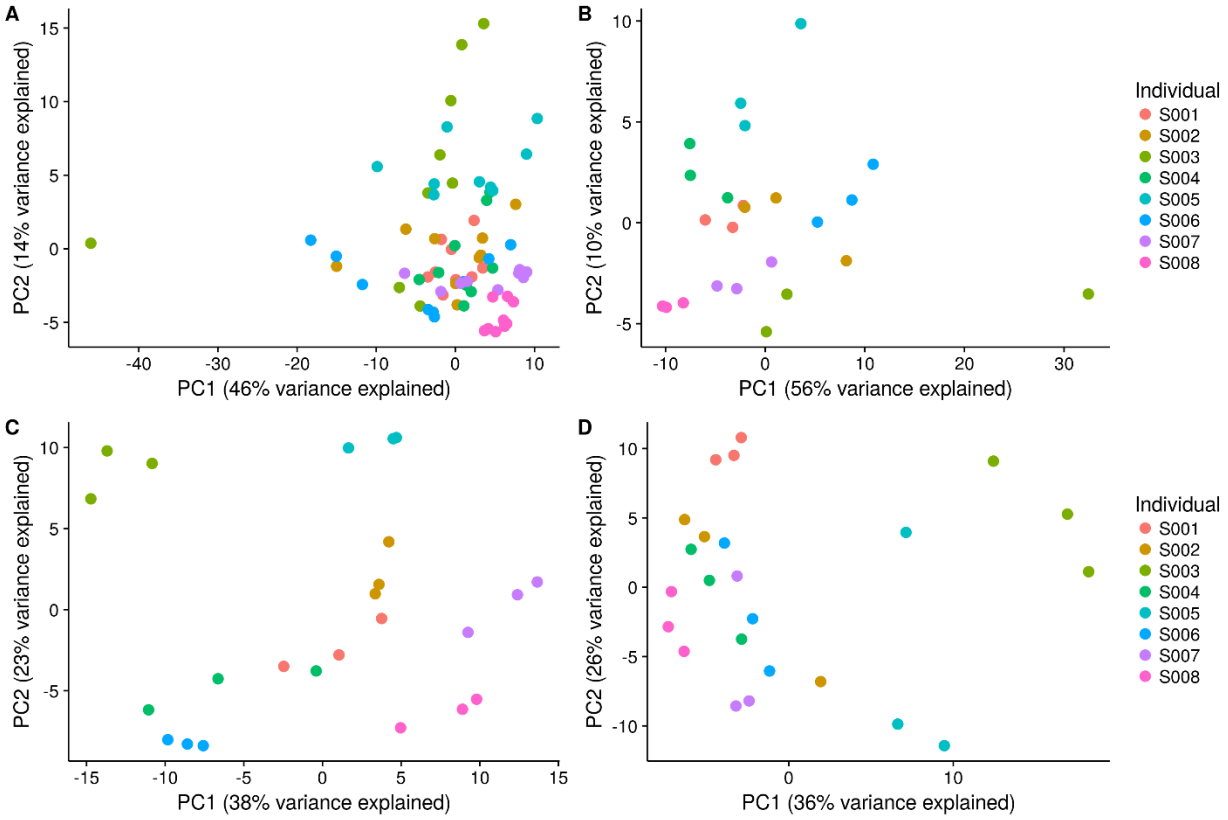

Figure 5. Principal component analysis of the nucleotide diversity of universal hidSkinPlex markers for each body site (threshold, ≥10x read depth). A) All samples regardless of body site. B) Toe web/ball of the foot (Fb). C) Palm of the non-dominant hand (Hp). D) Manubrium (Mb).

RMLR and 1NN classification was used to attribute skin microbiome samples to their respective individual donors using LOOCV. RMLR and 1NN were performed using skin microbiome profiles comprised of universal and non-universal markers at each read depth

threshold (i.e., 2x, 10x, 25x, 50x, 100x, 150x, 200x) (Figure 6). Classification accuracies (i.e., the percentage of samples classified correctly) were highest for Hp, ranging from 95.83-100% (average 97.92% ± 2.08 (SD)) using 98-207 (threshold 200x and 2x, respectively) universal markers (Table S3). Classification accuracies for Mb (threshold 200x and 25x/50x/150x, respectively) ranged from 70.83-95.83% (average 86.31% ± 6.94 (SD)). Classification accuracies calculated using enriched hidSkinPlex markers from the Hp and Mb were comparable and not significantly different than classification accuracies calculated using shotgun data [26] ($p = 1$ for Hp and Mb; Fisher's Exact Test). The hidSkinPlex enrichment successfully amplified common markers shared by all individuals on Fb, 37-188 markers (threshold 200x and 2x, respectively). The Fb results are substantially different from using shotgun sequencing data, where only 2-5 markers were common to individuals [26]. Classification accuracies for the Fb ranged from 54.17-83.33% (average 73.21% ± 7.51 (SD)). Another notable difference using targeted enrichment of common markers across body sites was the ability to classify microbiomes to their respective donor using all samples, when the body site was unknown to the classifier, in contrast to previous studies when the body site was known/assumed [9,24–26]. Classification accuracies for all samples ranged from 68.06-97.22% (average 87.60% ± 7.67 (SD)) using 17-183 markers (threshold 200x and 2x, respectively). RMLR and 1NN also were performed using non-universal markers at each threshold; however, average classification accuracies were lower for all body sites (Table S4). The only improvement using non-universal markers was an increase in classification accuracy up to 91.67% (threshold 10x) using 254 markers on Fb. To compare classification accuracies using targeted markers and shotgun data from the Fb, RMLR and 1NN were performed using shotgun data from the plantar heel (Ph), toenail (Tn), and toe web space (Tw) (body sites excluded from Schmedes et al. [26]) and were found to be significantly lower than classification

accuracies calculated using enriched hidSkinPlex markers ($p < 0.00001$; Fisher's Exact Test). The highest classification accuracy from the foot using shotgun data was 23% at the Tw body site.

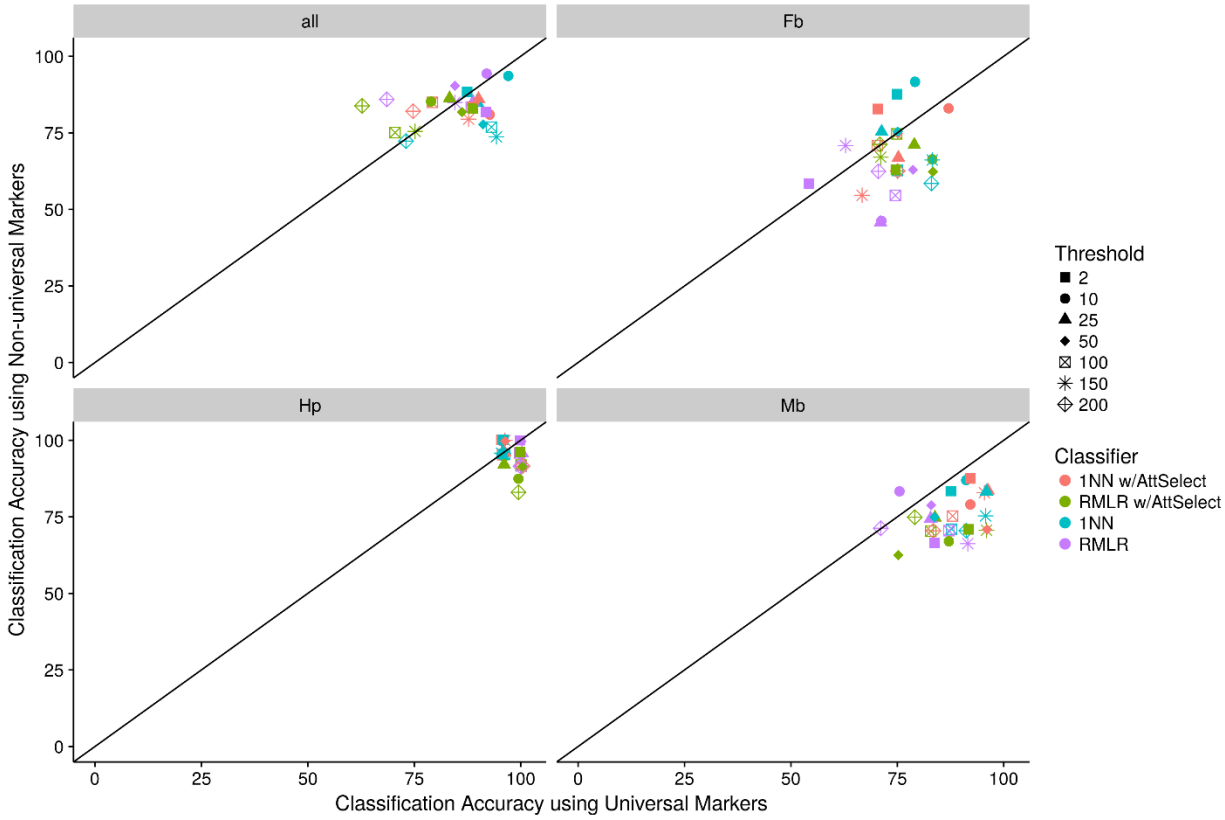

Figure 6. Comparison of skin microbiome classification accuracies using universal and non-universal hidSkinPlex markers. RMLR and 1NN, with and without attribute selection, were performed to attribute skin microbiomes to their respective individual host at each body site (i.e., all samples (all), toe web/ball of the foot (Fb), palm of non-dominant hand (Hp), and manubrium (Mb)).

Attribute selection (see Methods) was performed using LOOCV with RMLR and 1NN classification to determine if reduced subsets of hidSkinPlex markers produce comparable or increased classification accuracies (Figure 6). Additionally, attribute selection may allow for the selection of the most differentiating markers which may be better suited for microbiome profiling of particular body sites. hidSkinPlex marker subsets ranged in size from 8-20 (all), 15-31 (Fb), 38-

64 (Hp), and 13-43 markers (Mb) (Table S3). Classification accuracies using attribute selected markers were similar to accuracies using full sets of markers, a finding previously reported by Schmedes et al. [26] with shotgun metagenomic data. This finding also was observed when using non-universal markers (Table S4).

Propionibacterium acnes strain characterization

P. acnes has been shown to be a dominant skin flora with single-nucleotide variant (SNV) profiles [3], clade-specific marker phylogenies and pangenome gene presence/absence profiles that are stable over time [26]. To determine if *P. acnes* strains are more closely related at the individual level (i.e., regardless of body site) or more closely related at a particular body site for each individual, maximum-likelihood phylogenies were constructed using *P. acnes*-specific hidSkinPlex markers (> 65% of the hidSkinPlex panel) enriched in each body site to evaluate *P. acnes* strain-level variation across all body sites and individuals (Figures 7, S5-S7). If *P. acnes* strains are more closely related at the individual level, all nine samples from a particular individual would be more closely related and branch out from a common node, a pattern not observed in Figure 7. Only all nine samples for one individual form an individual-specific clade in the tree; however, samples from Mb and Hp from two additional individuals do form unique clades specific to those particular individuals. Instead, *P. acnes* strains tend to be more closely related if originating from the same individual and same body site, although some exceptions are evident (Figure 7). Unique individual-specific clades of *P. acnes* strains were most evident for Hp and Mb as compared to Fb (Figure S5-S7) and less diversity was observed between strains from different individuals in samples from Hp (Figure S6).

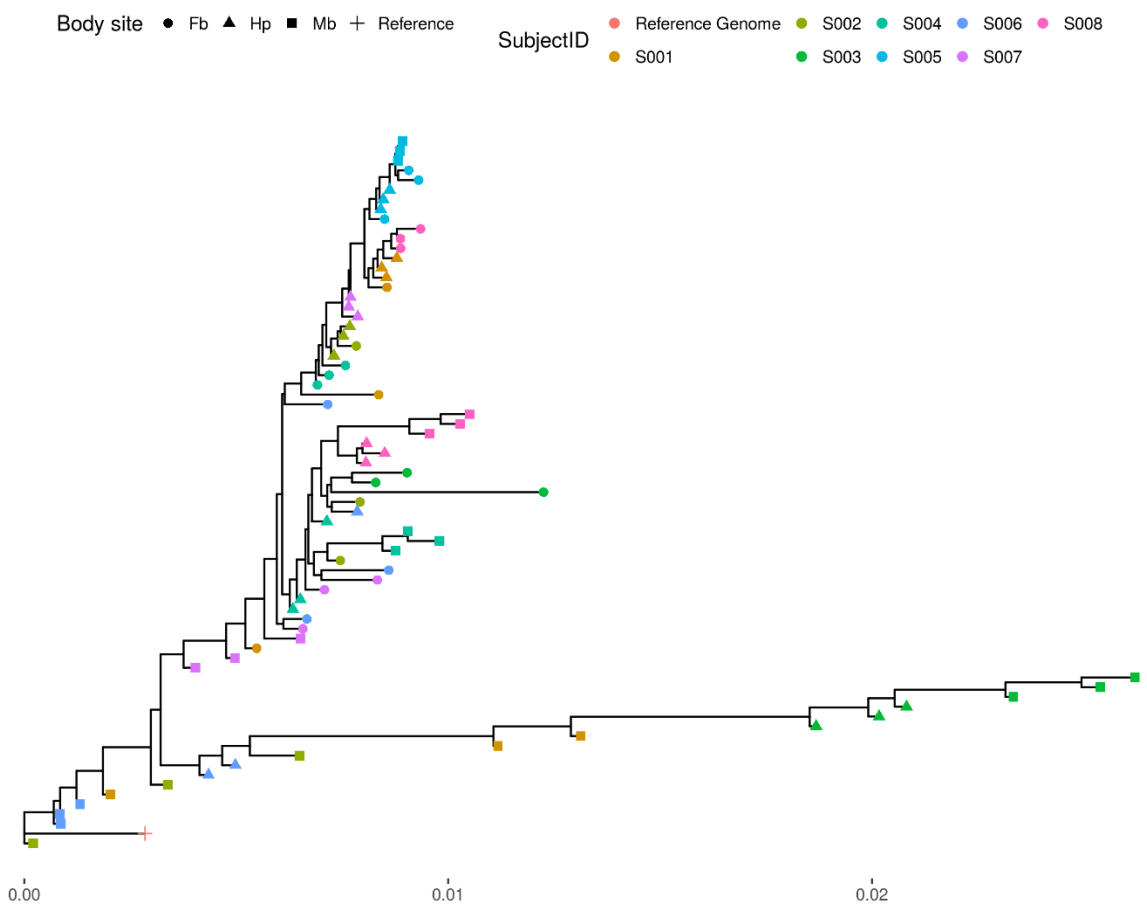

Figure 7. Maximum likelihood phylogeny of *Propionibacterium acnes* strains present in skin microbiomes from three skin body sites and eight individuals. The *P. acnes* phylogeny was constructed using all *P. acnes*-specific markers in the hidSkinPlex panel (n = 187).

Body site classification

The hidSkinPlex panel was developed with clade-specific markers selected for their ability to differentiate skin microbiome samples from different individuals. While the main purpose of the hidSkinPlex is for individual identification, body site identification was evaluated to determine if hidSkinPlex markers could serve a dual classification purpose. PCA of nucleotide diversities of non-universal hidSkinPlex markers showed clustering of skin microbiome samples from samples

collected from all three body sites, with greater variance observed across Fb than Hp and Mb, thus resolving Fb more so from Hp and Mb (Figure 8). RMLR and 1NN classification were performed, with and without attribute selection, as previously described, using skin microbiome profiles comprised of nucleotide diversities of non-universal hidSkinPlex markers to predict body site classification (Table S5). Body site classification was predicted with 69.44-86.11% accuracy (average 78.47% ± 4.16 (SD)) using 232-275 non-universal hidSkinPlex markers, respectively. Classification accuracies using 15-23 attribute selected markers were nearly identical (Table S5).

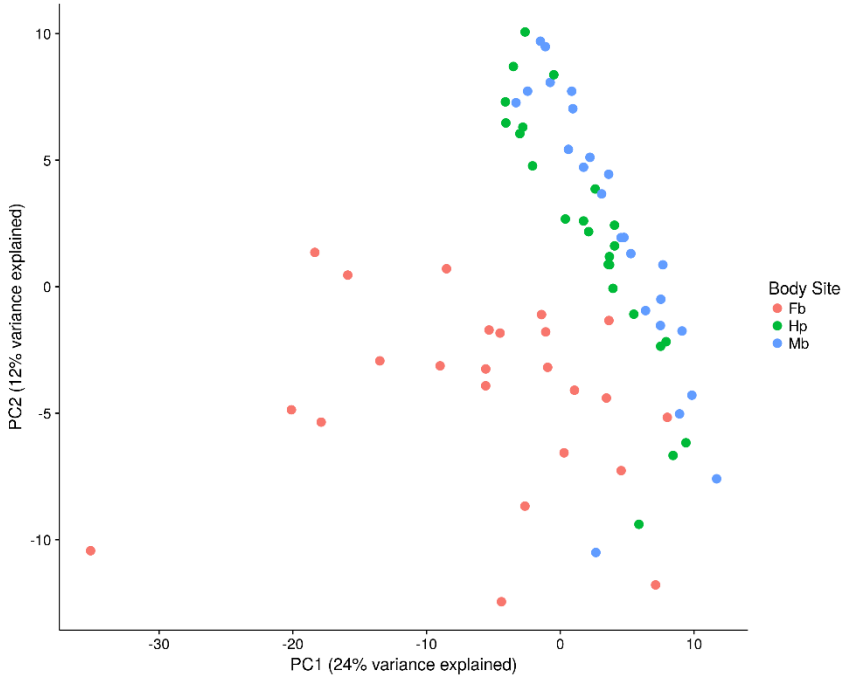

Figure 8. Principal component analysis of the nucleotide diversity of 261 non-universal hidSkinPlex markers for all body sites (threshold, ≥10x read depth).

hidSkinPlex profile coupled with human-specific STR and SNP profile

Skin microbiome profiling provides potential to generate additional identifying genetic data than human genetic profiles alone for human identification purposes. Given the higher copy

number of microbial cells to human cells, skin microbiome profiles may be used individually but also in conjunction with partial human DNA profiles for investigative purposes, especially from touched evidentiary items. To demonstrate this proof-of-concept, DNA extracts from female subject S001 from Fb, Hp, and Mb (n = 9), in addition to a buccal reference sample, were sequenced on the Illumina MiSeq FGx™ Forensic Genomics System using the Illumina ForenSeq™ panel (Primer Mix A). The recommended DNA input for the ForenSeq assay is 1 ng (5μL maximum input volume of template); however, DNA extracts from S001 (50 μL total volume) were low bio-mass samples with total DNA concentrations ranging from < 0.5 pg/μL for samples from the hand Hp and Mb and 56-86 pg/μL for samples from Fb. Thus all samples were below the ForenSeq optimum input recommendation. Only one of the nine samples yielded a full profile (Mb, replicate #3), while 8/9 samples yielded partial profiles ranging from 32% (Hp, replicate #1) to 99% (Mb, replicate #2) alleles detected (Figure 9). The lowest numbers of alleles detected were from samples collected from the hand (Hp), the same samples which were classified with 100% accuracy using the hidSkinPlex profile (Figure 6). Considering these skin samples were swabbed directly from the skin of the subject, similar trends (if not lower amounts) would likely be recovered from touch items in a forensic setting. The potentially more robust microbial profiles might be able to increase the strength of an association of a sample with a donor.

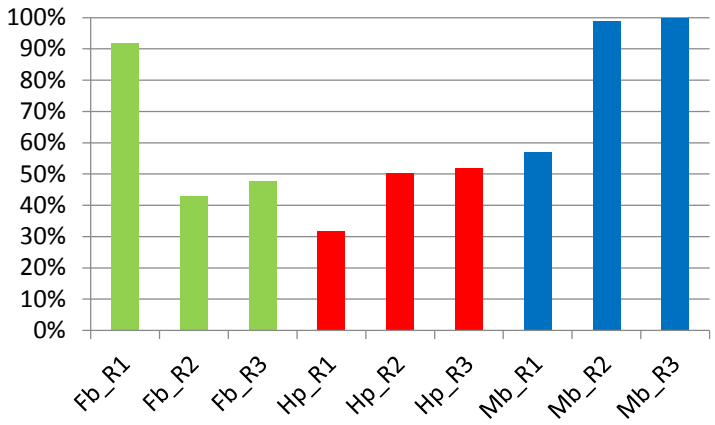

Figure 9. Percentage of ForenSeq STR/SNP alleles detected from skin swabs collected from subject S001 from the toe web/ball of the foot (Fb = green), palm of the non-dominant hand (Hp = red), and manubrium (Mb = blue). ForenSeq profiles were generated from skin swab samples collected from female subject S001 and compared to a buccal reference swab to determine of the percentage of ForenSeq STR and SNP alleles (n = 195) called from low-biomass skin swab samples.

Additional trace human alleles were detected from all nine skin swab samples from female subject S001. This observation in and of itself was not surprising, considering human skin comes into contact with touched objects and other people on a daily basis; however, in some cases the trace alleles were the major contributor (File S1). Several of these trace alleles were Y-chromosome STR alleles, potentially from 1-2 male donors (File S1). The majority of Y-STR alleles (n = 14 loci) were detected from skin samples collected from Fb, although alleles from the suspected male donors could be detected across all three body sites. One could presume these likely come from cohabitating male family member(s) such as a partner/spouse or other relatives. Future studies would need to be conducted to collect samples from cohabiting individuals to test the hypothesis that trace levels of human DNA, as well as shared microbial DNA, are prevalent on the surface of the skin for periods of time. Trace human and microbial DNA profiles might be able to determine close contact and frequency of contact between individuals, potentially assisting sexual assault investigations.

DISCUSSION

In this study, the hidSkinPlex, a novel targeted sequencing panel for skin microbiome profiling for forensic human identification, is described. The hidSkinPlex was developed to create a targeted enrichment solution to maximize marker detection and read depth for skin microbiome profiling. The hidSkinPlex is comprised of 286 bacterial (and phage) family-, genus-, species-, and subspecies-level markers previously selected by Schmedes et al. [26] by mining shotgun metagenomic datasets from skin microbiomes, sampled from 12 individuals over a 3 year period. These markers were deemed likely to be informative to differentiate microbiomes from individuals with a high degree of accuracy. The hidSkinPlex was designed to be coupled with Nextera XT library preparation to sequence on the Illumina MiSeq system. Three bacterial controls (i.e., *P. acnes* Strain SK137, *P. granulosum* D-34, and *R. dentocariosa* Strain M567) were used to evaluate the performance of the hidSkinPlex and yielded >85% - 100% amplification of expected markers (Figure 3). The hidSkinPlex was evaluated on skin swab samples collected from eight individuals and three body sites (i.e., Fb, Hp, and Mb). Amplification of hidSkinPlex markers was successful for all samples (n = 72), with amplification of 282/286 markers across all individuals and body sites (average 2,117x sequencing read depth), and 183 markers were common to all samples. Four markers from *Propionibacterium* phage P100 A, *Propionibacterium* phage P1 1, *Propionibacterium* phage PAD20, and *Propionibacterium* phage PAS50 were not detected in collected skin microbiomes samples, as well as the bacterial controls. Possible explanations for not observing the phages are they were not present, they failed to co-extract, or amplification failed due to primer design or PCR conditions. The latter explanation may be likely given *Propionibacterium* phages are prevalent in the skin microbiome [2,3]. Further evaluation will be needed to determine the cause of the absence of phage markers.

Bacterial contamination was observed in both the reagents blanks for the control sequencing portion of the study as well as in the swab blanks sequenced along with subject skin microbiome samples. In the control sequencing run all reads in the reagent blank were < 70x read depth; however, an average 71x ± 183 (SD) read depth was observed for the swab blanks sequenced with subject samples. Microbial contamination within DNA extraction kits and laboratory water has been observed and highlighted as cause for caution for microbiome and other low-abundance microbial studies [41,42]. Two of the dominant genera in the hidSkinPlex panel, *Propionibacterium* and *Corynebacterium*, have been previously reported as common contaminants in reagents [41]. The performance of the hidSkinPlex was assessed using a threshold of ≥ 70x read depth, to subtract reads from the reagent blank, to calculate true positives and negatives. However, since it was unknown what markers to expect or observe in each of the skin microbiome samples, a threshold was not used to remove reads. Given the high classification accuracies observed in this study (e.g., up to 92% (Fb) - 100% (Hp)), contamination likely did not significantly interfere with classification. In future studies, deeper analysis of these contaminant reads could be used to bioinformatically remove known contaminant reads from subject samples. Since bacterial contamination in laboratory reagents is a common issue, reagent and swab blanks should always be processed through the entire workflow and sequenced to identify any contaminants which may be present in reagents and consumables.

Classification accuracies using enriched clade-specific markers with 1NN classification for the Hp (up to 100%) and Mb (up to 96%) (Figure 6, Table S3) were both comparable and not significantly different ($p = 1$ for Hp and Mb; Fisher's Exact Test) than accuracies observed from shotgun metagenomic data, using clade-specific markers with 1NN [26]. Additionally, hidSkinPlex markers from Fb were successfully amplified and yielded classification accuracies up

to 92% using non-universal markers (Table S4). Individual classification accuracies using skin microbiomes from Fb were significantly higher ($p < 0.00001$; Fisher's Exact Test) using enriched hidSkinPlex markers, as opposed to markers from shotgun data which only yielded up to 23% classification accuracy for the toe web space (Tw) foot site [26]. The ability to classify skin microbiomes from the foot using a targeted enrichment method is significant since the foot harbors highly variable and low-abundant microbial communities [3], hindering classification using shotgun metagenomic data [26]. The foot is a forensically relevant skin site, and Goga [22] attempted to associate skin microbiome samples collected from shoe insoles with the correct owners' of the shoes using unsupervised methods. The hidSkinPlex with RMLR and 1NN offers a supervised approach to identify skin microbiomes sampled from the foot.

Enrichment of hidSkinPlex markers provides the capability to identify skin microbiomes from individuals when the body site is not known to the classifier with up to 97% accuracy using markers shared across Fb, Hp, and Mb (Figure 6, Table S3-S4) and provides the ability to identify the body site of origin of the skin microbiome sample with up to 86% accuracy (Table S5). Thus, the hidSkinPlex can serve a dual purpose, providing a method to not only identify individuals but also predict the body site origin of skin microbiome evidentiary samples. While the hidSkinPlex was not originally designed for body site classification, the addition of body site specific markers would likely yield higher body site classification accuracies. Further analyses of body site specific markers from shotgun metagenomic data would need to be performed to assess the utility of additional marker inclusion to the hidSkinPlex panel for body site identification capabilities.

P. acnes is a highly informative, forensically relevant target due its high abundance on all skin surfaces and stability of individual-specific strain-level profiles [3,26]. Oh et al. [3] previously reported that *P. acnes* strain and SNV profiles are individual-specific and are similar across body

sites. Schmedes et al. [26] described the stability of individual-specific *P. acnes* pangenome gene presence/absence profiles and *P. acnes* clade-specific phylogenies at individual body sites. Since > 65% of hidSkinPlex markers are from *P. acnes*, *P. acnes* strain diversity was assessed across all body sites to determine if strains are more closely related to individuals regardless of body site or more closely related to individuals at a specific body site. With few exceptions, *P. acnes* strains tend to be more closely related by individual and body site, in contrast to findings from Oh et al. [3] (Figure 7). However, additional analysis of *P. acnes* strain-specific SNPs identified in Oh et al. [3], outside the hidSkinPlex markers, would need to be performed to make a more appropriate comparison. The fact that these samples are associated within individuals across body sites, in some cases, may indicate samples collected from other body sites may be sufficient to identify an individual, even if the forensic sample is from an un-tested body site. To partially test this, classification was performed using all hidSkinPlex markers without conditioning on the body site and accuracies remained high (Figure 6, Table S3-S4). Due to the influence of *P. acnes* markers on individual classification at Hp and Mb (Figures S6-S7), additional *P. acnes*-strain specific SNP loci may be informative additions to the hidSkinPlex panel for improved individual identification capability, especially across multiple body sites.

Skin microbiome profiling serves as a potential tool to use in conjunction with low-biomass or degraded samples which fail to yield full human STR/SNP profiles of touched evidentiary items. In a small case study, skin microbiome profiles (hidSkinPlex) and human forensic profiles (Illumina ForenSeq panel A) were generated from the same DNA extracts sampled for one female study subject to assess each profile type generated from the same low-biomass samples. Few samples yielded complete or nearly-complete (92-99%) STR/SNP profiles from Fb (n = 1) and Mb (n = 2). Only partial profiles, 32-52% complete, were generated for all samples from the hand;

however, for these same samples using hidSkinPlex, profiles were able to be classified to their respective individual host with 100% accuracy, highlighting the potential microbiome profiles can provide, especially when used in conjunction with partial human STR/SNP profiles. These samples were collected directly from the skin, and not touched evidentiary items; touched samples would likely yield lower profile completeness. Multiple trace alleles were detected on all skin surfaces sampled from subject S001, including Y-STR alleles, with some alleles comprising the major contributor to the profile (File S1). Although, spurious alleles would be expected at low levels, likely from coming into contact with daily objects and surfaces touched by other individuals, detection of alleles common in multiple samples, and in some cases the major contributor, are likely to be due to frequent contact of subject S001, such as a spouse or family member(s). In fact, previous studies have demonstrated that microflora are more commonly shared among cohabitating family members and couples than with individuals from different households [6,43]. Indeed, Ross et al. [43] reported that microflora from the foot were more similar among couples than other body sites. Interestingly, the majority of Y-STR alleles, potentially from 1 male donor, were detected on the foot from S001. Future studies will address the level of trace human DNA shared by cohabiting couples and family members, as well as microbial DNA using the hidSkinPlex, to determine the potential of using foreign human and microbial DNA from persons as trace evidence.

CONCLUSION

In this study, the initial development and evaluation of the hidSkinPlex, a targeted sequencing panel for skin microbiome profiling for forensic human identification, are presented. Skin microbiome profiles generated using the hidSkinPlex from the foot, hand, and manubrium

were attributed to their respective individual host with up to 92% (Fb) - 100% (Hp) accuracy. Additionally, body site origin could be predicted with up to 86% accuracy. Future studies will assess the stability of skin microbiomes collected over varying time intervals, skin microbiome identification from touch samples coupled with human genetic profiles, and the degree of shared microbiome signatures between cohabitating couples and family members. Additional markers for the foot body site, likely from *Corynebacterium spp.* (a common genus colonizing the foot) and body site specific markers will be evaluated for inclusion into the hidSkinPlex. Further development of the hidSkinPlex will remove redundant markers (i.e., keep attribute selected markers) and identify the most differentiating regions within each marker in order to reduce amplicon size of these regions. Since the hidSkinPlex is not yet optimized, primer redesign and concentrations will be further evaluated to provide more uniform coverage and read depth across markers. Finally, additional analysis and statistical methods will be explored to develop analysis and interpretation guidelines for use of skin microbiome profiling in the forensic setting.

REFERENCES

[1] E.A. Grice, H.H. Kong, S. Conlan, C.B. Deming, J. Davis, A.C. Young, G.G. Bouffard, R.W. Blakesley, P.R. Murray, E.D. Green, M.L. Turner, J. a Segre, Topographical and temporal diversity of the human skin microbiome., Science. 324 (2009) 1190–1192. doi:10.1126/science.1171700.

[2] G.D. Hannigan, J.S. Meisel, A.S. Tyldsley, Q. Zheng, B.P. Hodkinson, A.J. Sanmiguel, S. Minot, F.D. Bushman, E.A. Grice, A. Grice, The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome, MBio. 6 (2015) e01578-15. doi:10.1128/mBio.01578-15.Editor.

[3] J. Oh, A.L. Byrd, M. Park, H.H. Kong, J.A. Segre, Temporal Stability of the Human Skin Microbiome, Cell. 165 (2016) 854–866. doi:10.1016/j.cell.2016.04.008.

[4] R. Blekhman, J.K. Goodrich, K. Huang, Q. Sun, R. Bukowski, J.T. Bell, T.D. Spector, A. Keinan, R.E. Ley, D. Gevers, A.G. Clark, Host genetic variation impacts microbiome composition across human body sites, Genome Biol. 16 (2015) 191. doi:10.1186/s13059-015-0759-1.

[5] T. Yatsunenko, F.E. Rey, M.J. Manary, I. Trehan, M.G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R.N. Baldassano, A.P. Anokhin, A.C. Heath, B. Warner, J. Reeder, J. Kuczynski, J.G. Caporaso, C.A. Lozupone, C. Lauber, J.C. Clemente, D. Knights, R. Knight, J.I. Gordon, Human gut microbiome viewed across age and geography., Nature. 486 (2012) 222–227. doi:10.1038/nature11053.

[6] S.J. Song, C. Lauber, E.K. Costello, C. a Lozupone, G. Humphrey, D. Berg-Lyons, J.G. Caporaso, D. Knights, J.C. Clemente, S. Nakielny, J.I. Gordon, N. Fierer, R. Knight, Cohabiting family members share microbiota with one another and with their dogs., Elife. 2 (2013) e00458. doi:10.7554/eLife.00458.

[7] H.H. Kong, J. Oh, C. Deming, S. Conlan, E.A. Grice, M.A. Beatson, E. Nomicos, E.C. Polley, H.D. Komarow, NISC Comparative Sequence Program, P.R. Murray, M.L. Turner, J.A. Segre, Temporal shifts in the skin microbiome associated with disease flare and treatment in children with atopic dermatitis, Genome Res. 22 (2012) 850–859. doi:10.1101/gr.131029.111.

[8] N. Fierer, M. Hamady, C.L. Lauber, R. Knight, The influence of sex, handedness, and washing on the diversity of hand surface bacteria., Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 17994–17999. doi:10.1073/pnas.0807920105.

[9] E. a. Franzosa, K. Huang, J.F. Meadow, D. Gevers, K.P. Lemon, B.J.M. Bohannan, C. Huttenhower, Identifying personal microbiomes using metagenomic codes, Proc. Natl. Acad. Sci. 112 (2015) E2930–E2938. doi:10.1073/pnas.1423854112.

[10] D. Tautz, Hypervariabflity of simple sequences as a general source for polymorphic DNA markers, Nucleic Acids Res. 17 (1989) 6463–6471. doi:10.1093/nar/17.16.6463.

[11] A. Edwards, A. Civitello, H.A. Hammond, C.T. Caskey, DNA typing and genetic mapping with trimeric and tetrameric tandem repeats., Am. J. Hum. Genet. 49 (1991) 746–756.

[12] A. Edwards, H. a. Hammond, L. Jin, C.T. Caskey, R. Chakraborty, Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups, Genomics. 12 (1992) 241–253. doi:10.1016/0888-7543(92)90371-X.

[13] P.J. Collins, L.K. Hennessy, C.S. Leibelt, R.K. Roby, D.J. Reeder, P.A. Foxall, Developmental Validation of a Single-tube Amplification of the 13 CODIS STR Loci, D2S1338, D19S433, and Amelogenin: The AmpFlSTR Identifiler PCR Amplificaiont Kit, J. Forensic Sci. 49 (2004) JFS2002195. doi:10.1016/j.fsigen.2016.10.016.

[14] B.E. Krenke, A. Tereba, S.J. Anderson, E. Buel, S. Culhane, C.J. Finis, C.S. Tomsey, J.M. Zachetti, A. Masibay, D.R. Rabbach, E. a Amiott, C.J. Sprecher, Validation of a 16-locus fluorescent multiplex system., J. Forensic Sci. 47 (2002) 773–85. http://www.ncbi.nlm.nih.gov/pubmed/20457027.

[15] S. Flores, J. Sun, J. King, B. Budowle, Internal validation of the GlobalFiler??? Express PCR Amplification Kit for the direct amplification of reference DNA samples on a high-throughput automated workflow, Forensic Sci. Int. Genet. 10 (2014) 33–39. doi:10.1016/j.fsigen.2014.01.005.

[16] M.G. Ensenberger, K.A. Lenz, L.K. Matthies, G.M. Hadinoto, J.E. Schienman, A.J. Przech,

M.W. Morganti, D.T. Renstrom, V.M. Baker, K.M. Gawrys, M. Hoogendoorn, C.R. Steffen, P. Martín, A. Alonso, H.R. Olson, C.J. Sprecher, D.R. Storts, Developmental validation of the PowerPlex® Fusion 6C System, Forensic Sci. Int. Genet. 21 (2016) 134–144. doi:10.1016/j.fsigen.2015.12.011.

[17] J.L. King, B.L. LaRue, N.M. Novroski, M. Stoljarova, S.B. Seo, X. Zeng, D.H. Warshauer, C.P. Davis, W. Parson, A. Sajantila, B. Budowle, High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq., Forensic Sci. Int. Genet. 12C (2014) 128–135. doi:10.1016/j.fsigen.2014.06.001.

[18] M.R. Wilson, J.A. DiZinno, D. Polanskey, J. Replogle, B. Budowle, Validation of mitochondrial DNA sequencing for forensic casework analysis, Int. J. Legal Med. 108 (1995) 68–74. doi:10.1007/BF01369907.

[19] M.M. Holland, T.J. Parsons, Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework, Forensic Sci. Rev. 11 (1999) 21–50.

[20] E. a Grice, H.H. Kong, G. Renaud, A.C. Young, G.G. Bouffard, R.W. Blakesley, T.G. Wolfsberg, M.L. Turner, J. a Segre, A diversity profile of the human skin microbiota., Genome Res. 18 (2008) 1043–1050. doi:10.1101/gr.075549.107.

[21] N. Fierer, C.L. Lauber, N. Zhou, D. McDonald, E.K. Costello, R. Knight, Forensic identification using skin bacterial communities., Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 6477–6481. doi:10.1073/pnas.1000162107.

[22] H. Goga, Comparison of bacterial DNA profiles of footwear insoles and soles of feet for the forensic discrimination of footwear owners, Int. J. Legal Med. 126 (2012) 815–823. doi:10.1007/s00414-012-0733-3.

[23] J.F. Meadow, A.E. Altrichter, J.L. Green, Mobile phones carry the personal microbiome of their owners., PeerJ. 2 (2014) e447. doi:10.7717/peerj.447.

[24] S. Lax, J.T. Hampton-Marcell, S.M. Gibbons, G.B. Colares, D. Smith, J. a Eisen, J. a Gilbert, Forensic analysis of the microbiome of phones and shoes, Microbiome. 3 (2015) 21. doi:10.1186/s40168-015-0082-9.

[25] D.W. Williams, G. Gibson, Individualization of pubic hair bacterial communities and the effects of storage time and temperature, Forensic Sci. Int. Genet. 26 (2017) 12–20. doi:10.1016/j.fsigen.2016.09.006.

[26] S.E. Schmedes, A.E. Woerner, B. Budowle, Forensic human identification using skin microbiomes, Appl. Environ. Microbiol. (2017) (In press).

[27] D.T. Truong, E.A. Franzosa, T.L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, N. Segata, MetaPhlAn2 for enhanced metagenomic taxonomic profiling, Nat. Methods. 12 (2015) 902–903. doi:10.1038/nmeth.3589.

[28] QIAGEN, QIAGEN Multiplex PCR Plus Handbook, https://www.qiagen.com/, 2016. https://www.qiagen.com/.

[29] J.L. King, F.R. Wendt, J. Sun, B. Budowle, STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems, Forensic Sci. Int. Genet. 29

(2017) 21–28. doi:10.1016/j.fsigen.2017.03.013.

[30] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.journal. 17 (2011) 10. doi:10.14806/ej.17.1.200.

[31] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools, Bioinformatics. 25 (2009) 2078–2079. doi:10.1093/bioinformatics/btp352.

[32] S. Nayfach, K.S. Pollard, Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria, bioRxiv. (2015) DOI:10.1101/031757. doi:10.1101/031757.

[33] E. Frank, M.A. Hall, I.H. Witten, Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," in: M. Kauffmann (Ed.), WEKA Work., Fourth Edi, 2016.

[34] S. Dorai-Raj, binom: Binomial Confidence Intervals for Several Parameterizations. R package version 1.1-1, (2014). https://cran.r-project.org/package=binom.

[35] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2009.

[36] C.O. Wilke, cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2". R package version 0.7.0, (2016). https://cran.r-project.org/package=cowplot.

[37] R.C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797. doi:10.1093/nar/gkh340.

[38] A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics. 30 (2014) 1312–1313. doi:10.1093/bioinformatics/btu033.

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

[40] G. Yu, D.K. Smith, H. Zhu, Y. Guan, T.T.Y. Lam, ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data, Methods Ecol. Evol. 8 (2017) 28–36. doi:10.1111/2041-210X.12628.

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

[42] M. Laurence, C. Hatzis, D.E. Brash, Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes., PLoS One. 9 (2014) e97876. doi:10.1371/journal.pone.0097876.

[43] A.A. Ross, A.C. Doxey, J.D. Neufeld, The Skin Microbiome of Cohabiting Couples, mSystems. 2 (2017) e00043-17. doi:10.1128/mSystems.00043-17.

ACKNOWLEDGEMENTS

We would like to acknowledge and thank Kameran Wong and Cydne Holt from Illumina, Inc. for their support and technical assistance with primer and multiplex design. We also would like to thank the subject volunteers for contributing skin swab samples for this study.

DATA AVAILABILITY

Sequence datasets can be found on the NCBI Sequence Read Archive under BioProject ID accession PRJNA398026. Custom perl and R scripts can be accessed at https://github.com/SESchmedes/hidSkinPlex.

COMPETING INTERESTS

Kathryn Stephens is employed by Verogen, Inc.

FUNDING

This project was supported by the National Institute of Justice, Award Number 2015-NE-BX-K006, the Texas Branch of the American Society for Microbiology, 2014 Eugene and Millicent Goldschmidt Graduate Student Award, and the Department of Defense, Award Number W911NF-16-0085.

SUPPLEMENTAL MATERIALS

Table S1. hidSkinPlex Markers

Marker	Clade	Level	Reference Marker Length (bp)	Amplicon Length (bp)
gi|260579795|ref|NZ_GG700815.1|:37946-39418	Corynebacterium_jeikeium	Species	1473	703
gi|260579796|ref|NZ_GG700816.1|:c282041-281199	Corynebacterium_jeikeium	Species	843	690
gi|311741741|ref|NZ_GL542877.1|:10765-11772	Corynebacterium_pseudogenitalium	Species	1008	688
gi|311741741|ref|NZ_GL542877.1|:c274839-274201	Corynebacterium_pseudogenitalium	Species	639	607
gi|311741741|ref|NZ_GL542877.1|:c32675-32286	Corynebacterium_pseudogenitalium	Species	390	384
gi|552779524|ref|NZ_KI515721.1|:c325186-324281	Corynebacterium_pseudogenitalium	Species	906	703
gi|512466269|ref|NZ_KE150404.1|:c2352553-2351375	Corynebacterium_sp_HFH0082	Species	1179	713
gi|552861940|ref|NZ_KI515759.1|:465825-466694	Corynebacterium_sp_KPL1818	Species	870	699
gi|552862639|ref|NZ_KI515762.1|:c2437-995	Corynebacterium_sp_KPL1818	Species	1443	700
gi|552839652|ref|NZ_KI515749.1|:427956-429677	Corynebacterium_sp_KPL1824	Species	1722	702
gi|255324262|ref|NZ_ACVP01000008.1|:c30911-29955	Corynebacterium_tuberculostearicum	Species	957	662
gi|255324262|ref|NZ_ACVP01000008.1|:c8199-6985	Corynebacterium_tuberculostearicum	Species	1215	715
gi|255324379|ref|NZ_ACVP01000012.1|:191207-192175	Corynebacterium_tuberculostearicum	Species	969	688
gi|255324379|ref|NZ_ACVP01000012.1|:c133969-133310	Corynebacterium_tuberculostearicum	Species	660	629
gi|255324379|ref|NZ_ACVP01000012.1|:c224628-223117	Corynebacterium_tuberculostearicum	Species	1512	703
gi|255324842|ref|NZ_ACVP01000019.1|:c614-294	Corynebacterium_tuberculostearicum	Species	321	266
gi|255324988|ref|NZ_ACVP01000023.1|:c144466-143744	Corynebacterium_tuberculostearicum	Species	723	678
gi|255324988|ref|NZ_ACVP01000023.1|:c170675-169767	Corynebacterium_tuberculostearicum	Species	909	700
gi|255325532|ref|NZ_ACVP01000028.1|:4502-4756	Corynebacterium_tuberculostearicum	Species	255	90
gi|255325617|ref|NZ_ACVP01000031.1|:c2144-699	Corynebacterium_tuberculostearicum	Species	1446	700
gi|552803646|ref|NZ_KI515731.1|:830184-831197	Corynebacterium_tuberculostearicum	Species	1014	688
gi|552803646|ref|NZ_KI515731.1|:c379068-378463	Corynebacterium_tuberculostearicum	Species	606	573
gi|552812292|ref|NZ_KI515735.1|:1411846-1412238	Corynebacterium_tuberculostearicum	Species	393	372
gi|552812292|ref|NZ_KI515735.1|:c1995797-1994775	Corynebacterium_tuberculostearicum	Species	1023	696
gi|552850245|ref|NZ_KI515751.1|:304189-304659	Corynebacterium_tuberculostearicum	Species	471	447
gi|552867507|ref|NZ_KI515768.1|:184972-186138	Corynebacterium_tuberculostearicum	Species	1167	700
gi|417931402|ref|NZ_AFUN01000007.1|:c97233-97075	GCF_000221145	Subspecies	159	126
gi|417932374|ref|NZ_AFUN01000032.1|:143771-144007	GCF_000221145	Subspecies	237	193
gi|417932959|ref|NZ_AFUN01000038.1|:225703-226005	GCF_000221145	Subspecies	303	293
gi|417933187|ref|NZ_AFUN01000043.1|:4929-5147	GCF_000221145	Subspecies	219	195
gi|335050656|ref|NZ_AFIK01000017.1|:c30516-30079	Propionibacteriaceae	Family	438	420
gi|335053539|ref|NZ_AFIL01000025.1|:23315-23623	Propionibacteriaceae	Family	309	295
gi|335055158|ref|NZ_AFIL01000073.1|:77143-77310	Propionibacteriaceae	Family	168	158
gi|342211239|ref|NZ_AFUK01000001.1|:c359834-359544	Propionibacteriaceae	Family	291	218
gi|355707189|ref|NZ_JH376566.1|:c170886-169537	Propionibacteriaceae	Family	1350	700
gi|552879811|ref|NZ_AXME01000001.1|:c2014536-2014075	Propionibacteriaceae	Family	462	357
gi|552891898|ref|NZ_AXMG01000001.1|:c1945194-1944973	Propionibacteriaceae	Family	222	211
gi|552896688|ref|NZ_AXMI01000003.1|:c72034-71849	Propionibacteriaceae	Family	186	170

gi|552904108|ref|NZ_KI518468.1|:464070-464315	Propionibacteriaceae	Family	246	237
gi|295129529|ref|NC_014039.1|:c1439020-1438442	Propionibacterium_acnes	Species	579	483
gi|335050281|ref|NZ_AFIK01000001.1|:c2940-1807	Propionibacterium_acnes	Species	1134	695
gi|335050542|ref|NZ_AFIK01000013.1|:c12739-12119	Propionibacterium_acnes	Species	621	590
gi|335050601|ref|NZ_AFIK01000014.1|:3050-3691	Propionibacterium_acnes	Species	642	580
gi|335050601|ref|NZ_AFIK01000014.1|:315-1133	Propionibacterium_acnes	Species	819	697
gi|335050697|ref|NZ_AFIK01000020.1|:c12439-12299	Propionibacterium_acnes	Species	141	80
gi|335050749|ref|NZ_AFIK01000022.1|:c35390-34998	Propionibacterium_acnes	Species	393	371
gi|335050796|ref|NZ_AFIK01000023.1|:c3954-3715	Propionibacterium_acnes	Species	240	229
gi|335051081|ref|NZ_AFIK01000036.1|:c1716-193	Propionibacterium_acnes	Species	1524	701
gi|335051327|ref|NZ_AFIK01000049.1|:8242-9012	Propionibacterium_acnes	Species	771	714
gi|335051382|ref|NZ_AFIK01000053.1|:c36245-34977	Propionibacterium_acnes	Species	1269	698
gi|335051382|ref|NZ_AFIK01000053.1|:c47134-46805	Propionibacterium_acnes	Species	330	244
gi|335051798|ref|NZ_AFIK01000065.1|:c4330-4001	Propionibacterium_acnes	Species	330	316
gi|335052272|ref|NZ_AFIK01000082.1|:c111360-110575	Propionibacterium_acnes	Species	786	690
gi|335052413|ref|NZ_AFIK01000085.1|:c27721-27527	Propionibacterium_acnes	Species	195	169
gi|335053104|ref|NZ_AFIL01000010.1|:c33862-32210	Propionibacterium_acnes	Species	1653	703
gi|335053685|ref|NZ_AFIL01000030.1|:c57253-57113	Propionibacterium_acnes	Species	141	103
gi|335053685|ref|NZ_AFIL01000030.1|:c58004-57372	Propionibacterium_acnes	Species	633	596
gi|335053761|ref|NZ_AFIL01000031.1|:46041-46637	Propionibacterium_acnes	Species	597	551
gi|335054110|ref|NZ_AFIL01000040.1|:4048-4263	Propionibacterium_acnes	Species	216	184
gi|335054576|ref|NZ_AFIL01000053.1|:7953-8144	Propionibacterium_acnes	Species	192	137
gi|335054619|ref|NZ_AFIL01000056.1|:14685-15386	Propionibacterium_acnes	Species	702	664
gi|335054657|ref|NZ_AFIL01000058.1|:28786-29034	Propionibacterium_acnes	Species	249	188
gi|335054657|ref|NZ_AFIL01000058.1|:c4236-3517	Propionibacterium_acnes	Species	720	683
gi|335054695|ref|NZ_AFIL01000059.1|:c28497-27568	Propionibacterium_acnes	Species	930	705
gi|335055047|ref|NZ_AFIL01000069.1|:c9632-8838	Propionibacterium_acnes	Species	795	681
gi|335055061|ref|NZ_AFIL01000070.1|:3643-4386	Propionibacterium_acnes	Species	744	700
gi|342211239|ref|NZ_AFUK01000001.1|:1588290-1589009	Propionibacterium_acnes	Species	720	689
gi|342211239|ref|NZ_AFUK01000001.1|:1828645-1829349	Propionibacterium_acnes	Species	705	695
gi|342211239|ref|NZ_AFUK01000001.1|:1851240-1852028	Propionibacterium_acnes	Species	789	704
gi|342211239|ref|NZ_AFUK01000001.1|:2001142-2001459	Propionibacterium_acnes	Species	318	263
gi|342211239|ref|NZ_AFUK01000001.1|:2069064-2069282	Propionibacterium_acnes	Species	219	186
gi|342211239|ref|NZ_AFUK01000001.1|:527724-528653	Propionibacterium_acnes	Species	930	699
gi|342211239|ref|NZ_AFUK01000001.1|:535213-535428	Propionibacterium_acnes	Species	216	160
gi|342211239|ref|NZ_AFUK01000001.1|:593413-594699	Propionibacterium_acnes	Species	1287	695
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	Propionibacterium_acnes	Species	1323	700
gi|342211239|ref|NZ_AFUK01000001.1|:c1255510-1255055	Propionibacterium_acnes	Species	456	442
gi|342211239|ref|NZ_AFUK01000001.1|:c1376325-1376110	Propionibacterium_acnes	Species	216	215
gi|342211239|ref|NZ_AFUK01000001.1|:c1579497-1578787	Propionibacterium_acnes	Species	711	667
gi|342211239|ref|NZ_AFUK01000001.1|:c1715790-1715233	Propionibacterium_acnes	Species	558	515
gi|342211239|ref|NZ_AFUK01000001.1|:c1845075-1844710	Propionibacterium_acnes	Species	366	308
gi|342211239|ref|NZ_AFUK01000001.1|:c1936798-1936352	Propionibacterium_acnes	Species	447	402
gi|342211239|ref|NZ_AFUK01000001.1|:c395948-395412	Propionibacterium_acnes	Species	537	467
gi|355707189|ref|NZ_JH376566.1|:1026577-1027557	Propionibacterium_acnes	Species	981	696
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	Propionibacterium_acnes	Species	1278	697
gi|355707189|ref|NZ_JH376566.1|:1105369-1105965	Propionibacterium_acnes	Species	597	576

gi|355707189|ref|NZ_JH376566.1|:326756-326986	Propionibacterium_acnes	Species	231	107
gi|355707189|ref|NZ_JH376566.1|:507019-507612	Propionibacterium_acnes	Species	594	543
gi|355707189|ref|NZ_JH376566.1|:882552-883256	Propionibacterium_acnes	Species	705	690
gi|355707384|ref|NZ_JH376567.1|:190789-191232	Propionibacterium_acnes	Species	444	430
gi|355707384|ref|NZ_JH376567.1|:251291-251998	Propionibacterium_acnes	Species	708	669
gi|355707384|ref|NZ_JH376567.1|:592116-592328	Propionibacterium_acnes	Species	213	209
gi|355707384|ref|NZ_JH376567.1|:598376-599065	Propionibacterium_acnes	Species	690	639
gi|355707384|ref|NZ_JH376567.1|:621102-621467	Propionibacterium_acnes	Species	366	360
gi|355707384|ref|NZ_JH376567.1|:90374-91453	Propionibacterium_acnes	Species	1080	694
gi|355707384|ref|NZ_JH376567.1|:c379886-379035	Propionibacterium_acnes	Species	852	721
gi|355707384|ref|NZ_JH376567.1|:c388018-387605	Propionibacterium_acnes	Species	414	348
gi|355707384|ref|NZ_JH376567.1|:c400475-400284	Propionibacterium_acnes	Species	192	161
gi|355708280|ref|NZ_JH376568.1|:c185858-185226	Propionibacterium_acnes	Species	633	576
gi|355708280|ref|NZ_JH376568.1|:c255689-255105	Propionibacterium_acnes	Species	585	491
gi|355708440|ref|NZ_JH376569.1|:c80380-79448	Propionibacterium_acnes	Species	933	710
gi|365961730|ref|NC_016511.1|:2485446-2486162	Propionibacterium_acnes	Species	717	677
gi|386069650|ref|NC_017550.1|:821046-821639	Propionibacterium_acnes	Species	594	441
gi|387502364|ref|NC_017535.1|:c1339878-1339075	Propionibacterium_acnes	Species	804	675
gi|407934369|ref|NC_018707.1|:c1315368-1314979	Propionibacterium_acnes	Species	390	380
gi|417929021|ref|NZ_AFUM01000003.1|:557611-558279	Propionibacterium_acnes	Species	669	586
gi|422385765|ref|NZ_GL878448.1|:c80834-80607	Propionibacterium_acnes	Species	228	213
gi|422386402|ref|NZ_GL878455.1|:c805995-805537	Propionibacterium_acnes	Species	459	413
gi|422386402|ref|NZ_GL878455.1|:c812899-812252	Propionibacterium_acnes	Species	648	646
gi|422388755|ref|NZ_GL878472.1|:c178957-178325	Propionibacterium_acnes	Species	633	572
gi|422392301|ref|NZ_GL883048.1|:64439-65218	Propionibacterium_acnes	Species	780	687
gi|422423570|ref|NZ_GL384259.1|:c300859-299957	Propionibacterium_acnes	Species	903	703
gi|422434141|ref|NZ_GL384222.1|:86635-86934	Propionibacterium_acnes	Species	300	284
gi|422436532|ref|NZ_GL384462.1|:c297812-297150	Propionibacterium_acnes	Species	663	608
gi|422439172|ref|NZ_GL384485.1|:c80610-80086	Propionibacterium_acnes	Species	525	502
gi|422482616|ref|NZ_GL383714.1|:170052-170369	Propionibacterium_acnes	Species	318	259
gi|422496709|ref|NZ_GL383802.1|:56803-56916	Propionibacterium_acnes	Species	114	95
gi|422499020|ref|NZ_GL383811.1|:10443-11039	Propionibacterium_acnes	Species	597	522
gi|422500804|ref|NZ_GL383759.1|:c166532-166311	Propionibacterium_acnes	Species	222	200
gi|422511741|ref|NZ_GL383929.1|:146431-146739	Propionibacterium_acnes	Species	309	266
gi|422512600|ref|NZ_GL383846.1|:26161-26922	Propionibacterium_acnes	Species	762	699
gi|422538210|ref|NZ_GL384610.1|:c285619-284684	Propionibacterium_acnes	Species	936	653
gi|422539030|ref|NZ_GL384611.1|:c783227-783054	Propionibacterium_acnes	Species	174	122
gi|422547321|ref|NZ_GL383459.1|:130129-130737	Propionibacterium_acnes	Species	609	585
gi|422552858|ref|NZ_GL383469.1|:c216727-215501	Propionibacterium_acnes	Species	1227	701
gi|482889214|ref|NC_021085.1|:654926-655153	Propionibacterium_acnes	Species	228	205
gi|552875787|ref|NZ_KI515684.1|:459339-460115	Propionibacterium_acnes	Species	777	688
gi|552875787|ref|NZ_KI515684.1|:489358-490317	Propionibacterium_acnes	Species	960	705
gi|552875787|ref|NZ_KI515684.1|:c325537-325361	Propionibacterium_acnes	Species	177	146
gi|552875787|ref|NZ_KI515684.1|:c44215-43715	Propionibacterium_acnes	Species	501	467
gi|552875787|ref|NZ_KI515684.1|:c488989-488798	Propionibacterium_acnes	Species	192	153
gi|552875787|ref|NZ_KI515684.1|:c584270-583890	Propionibacterium_acnes	Species	381	296
gi|552875787|ref|NZ_KI515684.1|:c96934-96368	Propionibacterium_acnes	Species	567	470

gi|552876418|ref|NZ_KI515685.1|:1081256-1081411	Propionibacterium_acnes	Species	156	74
gi|552876418|ref|NZ_KI515685.1|:133418-133666	Propionibacterium_acnes	Species	249	219
gi|552876418|ref|NZ_KI515685.1|:187493-188140	Propionibacterium_acnes	Species	648	609
gi|552876418|ref|NZ_KI515685.1|:225601-226386	Propionibacterium_acnes	Species	786	684
gi|552876418|ref|NZ_KI515685.1|:339623-340705	Propionibacterium_acnes	Species	1083	692
gi|552876418|ref|NZ_KI515685.1|:36713-37258	Propionibacterium_acnes	Species	546	546
gi|552876418|ref|NZ_KI515685.1|:432422-433465	Propionibacterium_acnes	Species	1044	708
gi|552876418|ref|NZ_KI515685.1|:546580-547218	Propionibacterium_acnes	Species	639	569
gi|552876418|ref|NZ_KI515685.1|:656232-656693	Propionibacterium_acnes	Species	462	432
gi|552876418|ref|NZ_KI515685.1|:910-1341	Propionibacterium_acnes	Species	432	402
gi|552876418|ref|NZ_KI515685.1|:c1014617-1014117	Propionibacterium_acnes	Species	501	493
gi|552876418|ref|NZ_KI515685.1|:c1032381-1030873	Propionibacterium_acnes	Species	1509	696
gi|552876418|ref|NZ_KI515685.1|:c157510-157292	Propionibacterium_acnes	Species	219	201
gi|552876418|ref|NZ_KI515685.1|:c184358-183951	Propionibacterium_acnes	Species	408	354
gi|552876418|ref|NZ_KI515685.1|:c713438-713010	Propionibacterium_acnes	Species	429	406
gi|552876418|ref|NZ_KI515685.1|:c727842-726979	Propionibacterium_acnes	Species	864	707
gi|552876418|ref|NZ_KI515685.1|:c743399-743001	Propionibacterium_acnes	Species	399	392
gi|552876418|ref|NZ_KI515685.1|:c849089-848304	Propionibacterium_acnes	Species	786	703
gi|552876418|ref|NZ_KI515685.1|:c931935-931327	Propionibacterium_acnes	Species	609	538
gi|552876815|ref|NZ_KI515686.1|:323579-324514	Propionibacterium_acnes	Species	936	700
gi|552876815|ref|NZ_KI515686.1|:613740-614315	Propionibacterium_acnes	Species	576	442
gi|552876815|ref|NZ_KI515686.1|:c104786-104448	Propionibacterium_acnes	Species	339	267
gi|552876815|ref|NZ_KI515686.1|:c200743-199319	Propionibacterium_acnes	Species	1425	699
gi|552876815|ref|NZ_KI515686.1|:c50594-49899	Propionibacterium_acnes	Species	696	618
gi|552876815|ref|NZ_KI515686.1|:c586091-585333	Propionibacterium_acnes	Species	759	653
gi|552876815|ref|NZ_KI515686.1|:c642879-642748	Propionibacterium_acnes	Species	132	126
gi|552879811|ref|NZ_AXME01000001.1|:1088727-1089377	Propionibacterium_acnes	Species	651	550
gi|552879811|ref|NZ_AXME01000001.1|:1128888-1129136	Propionibacterium_acnes	Species	249	243
gi|552879811|ref|NZ_AXME01000001.1|:1146402-1146932	Propionibacterium_acnes	Species	531	529
gi|552879811|ref|NZ_AXME01000001.1|:1265476-1266570	Propionibacterium_acnes	Species	1095	691
gi|552879811|ref|NZ_AXME01000001.1|:1286960-1287442	Propionibacterium_acnes	Species	483	417
gi|552879811|ref|NZ_AXME01000001.1|:1327950-1328573	Propionibacterium_acnes	Species	624	619
gi|552879811|ref|NZ_AXME01000001.1|:287543-287779	Propionibacterium_acnes	Species	237	211
gi|552879811|ref|NZ_AXME01000001.1|:368977-369813	Propionibacterium_acnes	Species	837	697
gi|552879811|ref|NZ_AXME01000001.1|:40840-41742	Propionibacterium_acnes	Species	903	696
gi|552879811|ref|NZ_AXME01000001.1|:49241-49654	Propionibacterium_acnes	Species	414	404
gi|552879811|ref|NZ_AXME01000001.1|:587256-587825	Propionibacterium_acnes	Species	570	564
gi|552879811|ref|NZ_AXME01000001.1|:702826-703131	Propionibacterium_acnes	Species	306	252
gi|552879811|ref|NZ_AXME01000001.1|:97330-98208	Propionibacterium_acnes	Species	879	700
gi|552879811|ref|NZ_AXME01000001.1|:c1552174-1551533	Propionibacterium_acnes	Species	642	480
gi|552879811|ref|NZ_AXME01000001.1|:c1599141-1598893	Propionibacterium_acnes	Species	249	159
gi|552879811|ref|NZ_AXME01000001.1|:c1651715-1651248	Propionibacterium_acnes	Species	468	461
gi|552879811|ref|NZ_AXME01000001.1|:c1657647-1657093	Propionibacterium_acnes	Species	555	539
gi|552879811|ref|NZ_AXME01000001.1|:c2135959-2134715	Propionibacterium_acnes	Species	1245	698
gi|552879811|ref|NZ_AXME01000001.1|:c2447430-2446870	Propionibacterium_acnes	Species	561	476
gi|552879811|ref|NZ_AXME01000001.1|:c550719-550297	Propionibacterium_acnes	Species	423	378
gi|552891898|ref|NZ_AXMG01000001.1|:1150303-1151070	Propionibacterium_acnes	Species	768	694

gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	Propionibacterium_acnes	Species	621	589
gi|552891898|ref|NZ_AXMG01000001.1|:1234202-1234792	Propionibacterium_acnes	Species	591	548
gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	Propionibacterium_acnes	Species	252	213
gi|552891898|ref|NZ_AXMG01000001.1|:1877095-1877379	Propionibacterium_acnes	Species	285	238
gi|552891898|ref|NZ_AXMG01000001.1|:2120985-2121719	Propionibacterium_acnes	Species	735	685
gi|552891898|ref|NZ_AXMG01000001.1|:315632-315934	Propionibacterium_acnes	Species	303	293
gi|552891898|ref|NZ_AXMG01000001.1|:536557-537231	Propionibacterium_acnes	Species	675	619
gi|552891898|ref|NZ_AXMG01000001.1|:592123-592665	Propionibacterium_acnes	Species	543	526
gi|552891898|ref|NZ_AXMG01000001.1|:793445-793843	Propionibacterium_acnes	Species	399	389
gi|552891898|ref|NZ_AXMG01000001.1|:834824-835255	Propionibacterium_acnes	Species	432	425
gi|552891898|ref|NZ_AXMG01000001.1|:99114-99290	Propionibacterium_acnes	Species	177	165
gi|552891898|ref|NZ_AXMG01000001.1|:c1328090-1327596	Propionibacterium_acnes	Species	495	411
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	Propionibacterium_acnes	Species	603	601
gi|552891898|ref|NZ_AXMG01000001.1|:c1460921-1460529	Propionibacterium_acnes	Species	393	382
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	Propionibacterium_acnes	Species	528	515
gi|552891898|ref|NZ_AXMG01000001.1|:c2312839-2311925	Propionibacterium_acnes	Species	915	705
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	Propionibacterium_acnes	Species	399	381
gi|552891898|ref|NZ_AXMG01000001.1|:c2429318-2428110	Propionibacterium_acnes	Species	1209	697
gi|552895565|ref|NZ_AXMI01000001.1|:619555-620031	Propionibacterium_acnes	Species	477	466
gi|552895565|ref|NZ_AXMI01000001.1|:c101377-100163	Propionibacterium_acnes	Species	1215	697
gi|552895565|ref|NZ_AXMI01000001.1|:c14352-13837	Propionibacterium_acnes	Species	516	468
gi|552895565|ref|NZ_AXMI01000001.1|:c282323-281691	Propionibacterium_acnes	Species	633	607
gi|552895565|ref|NZ_AXMI01000001.1|:c29469-28930	Propionibacterium_acnes	Species	540	534
gi|552895565|ref|NZ_AXMI01000001.1|:c306684-306040	Propionibacterium_acnes	Species	645	598
gi|552895565|ref|NZ_AXMI01000001.1|:c325088-324501	Propionibacterium_acnes	Species	588	587
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	Propionibacterium_acnes	Species	1116	696
gi|552895565|ref|NZ_AXMI01000001.1|:c94830-94675	Propionibacterium_acnes	Species	156	108
gi|552896371|ref|NZ_AXMI01000002.1|:319095-319601	Propionibacterium_acnes	Species	507	467
gi|552896371|ref|NZ_AXMI01000002.1|:525312-525770	Propionibacterium_acnes	Species	459	447
gi|552896371|ref|NZ_AXMI01000002.1|:638332-638937	Propionibacterium_acnes	Species	606	556
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	Propionibacterium_acnes	Species	600	567
gi|552896371|ref|NZ_AXMI01000002.1|:721564-722400	Propionibacterium_acnes	Species	837	714
gi|552896371|ref|NZ_AXMI01000002.1|:837080-837400	Propionibacterium_acnes	Species	321	309
gi|552896371|ref|NZ_AXMI01000002.1|:c247178-246402	Propionibacterium_acnes	Species	777	698
gi|552896371|ref|NZ_AXMI01000002.1|:c671938-670697	Propionibacterium_acnes	Species	1242	708
gi|552896371|ref|NZ_AXMI01000002.1|:c872629-871631	Propionibacterium_acnes	Species	999	705
gi|552896688|ref|NZ_AXMI01000003.1|:232201-232740	Propionibacterium_acnes	Species	540	491
gi|552896688|ref|NZ_AXMI01000003.1|:c38494-37955	Propionibacterium_acnes	Species	540	486
gi|552897201|ref|NZ_AXMI01000004.1|:13568-14401	Propionibacterium_acnes	Species	834	700
gi|552897201|ref|NZ_AXMI01000004.1|:48085-48816	Propionibacterium_acnes	Species	732	679
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	Propionibacterium_acnes	Species	813	701
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	Propionibacterium_acnes	Species	555	530
gi|552897201|ref|NZ_AXMI01000004.1|:c577292-575922	Propionibacterium_acnes	Species	1371	700
gi|552897201|ref|NZ_AXMI01000004.1|:c732370-731744	Propionibacterium_acnes	Species	627	621
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	Propionibacterium_acnes	Species	447	384
gi|552902020|ref|NZ_AXMK01000001.1|:c1625038-1624022	Propionibacterium_acnes	Species	1017	704
gi|552902190|ref|NZ_AXML01000004.1|:c579659-578172	Propionibacterium_acnes	Species	1488	697

gi|544671929|ref|NZ_AOSS01000350.1|:c10780-9908	Propionibacterium_granulosum	Species	873	704
gi|544672317|ref|NZ_AOST01000022.1|:66190-68115	Propionibacterium_granulosum	Species	1926	703
gi|550735774|gb|AXMM01000002.1|:c751774-751298	Propionibacterium_granulosum	Species	477	463
gi|395203061|ref|NZ_AFAM01000001.1|:c244616-243831	Propionibacterium_humerusii	Species	786	698
gi|395203061|ref|NZ_AFAM01000001.1|:c260639-259980	Propionibacterium_humerusii	Species	660	658
gi|395203061|ref|NZ_AFAM01000001.1|:c312862-312554	Propionibacterium_humerusii	Species	309	288
gi|395203061|ref|NZ_AFAM01000001.1|:c34216-33161	Propionibacterium_humerusii	Species	1056	708
gi|395203469|ref|NZ_AFAM01000002.1|:37393-37605	Propionibacterium_humerusii	Species	213	208
gi|395203690|ref|NZ_AFAM01000005.1|:7982-10204	Propionibacterium_humerusii	Species	2223	705
gi|395203690|ref|NZ_AFAM01000005.1|:c111259-111038	Propionibacterium_humerusii	Species	222	193
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	Propionibacterium_humerusii	Species	126	90
gi|395203852|ref|NZ_AFAM01000006.1|:193159-193779	Propionibacterium_humerusii	Species	621	555
gi|395203852|ref|NZ_AFAM01000006.1|:75953-76378	Propionibacterium_humerusii	Species	426	361
gi|395203852|ref|NZ_AFAM01000006.1|:c137365-136916	Propionibacterium_humerusii	Species	450	431
gi|395203852|ref|NZ_AFAM01000006.1|:c75652-75533	Propionibacterium_humerusii	Species	120	310
gi|395204147|ref|NZ_AFAM01000008.1|:231579-231755	Propionibacterium_humerusii	Species	177	146
gi|395204147|ref|NZ_AFAM01000008.1|:c192705-192466	Propionibacterium_humerusii	Species	240	227
gi|395204147|ref|NZ_AFAM01000008.1|:c721415-721191	Propionibacterium_humerusii	Species	225	195
gi|395205131|ref|NZ_AFAM01000014.1|:c59116-58358	Propionibacterium_humerusii	Species	759	693
gi|395205131|ref|NZ_AFAM01000014.1|:c69464-69276	Propionibacterium_humerusii	Species	189	84
gi|395205346|ref|NZ_AFAM01000017.1|:12091-12363	Propionibacterium_humerusii	Species	273	208
gi|395205346|ref|NZ_AFAM01000017.1|:477016-477147	Propionibacterium_humerusii	Species	132	101
gi|395205346|ref|NZ_AFAM01000017.1|:c111452-110940	Propionibacterium_humerusii	Species	513	509
gi|395205346|ref|NZ_AFAM01000017.1|:c304806-304684	Propionibacterium_humerusii	Species	123	83
gi|395205346|ref|NZ_AFAM01000017.1|:c43269-42787	Propionibacterium_humerusii	Species	483	405
gi|395205346|ref|NZ_AFAM01000017.1|:c476952-476512	Propionibacterium_humerusii	Species	441	382
gi|395205346|ref|NZ_AFAM01000017.1|:c655204-654380	Propionibacterium_humerusii	Species	825	713
gi|395206111|ref|NZ_AFAM01000018.1|:111525-111779	Propionibacterium_humerusii	Species	255	210
gi|395206111|ref|NZ_AFAM01000018.1|:226375-226509	Propionibacterium_humerusii	Species	135	80
gi|395206455|ref|NZ_AFAM01000020.1|:c4555-4424	Propionibacterium_humerusii	Species	132	119
GeneID:13826912	Propionibacterium_phage_P100_A	Species	393	391
GeneID:13827106	Propionibacterium_phage_P1_1	Species	198	151
GeneID:10498655	Propionibacterium_phage_PAD20	Species	741	709
GeneID:10498608	Propionibacterium_phage_PAS50	Species	663	642
gi|335052938|ref|NZ_AFIL01000004.1|:4461-4578	Propionibacterium_sp_434_HC2	Species	118	96
gi|335053104|ref|NZ_AFIL01000010.1|:c43071-42837	Propionibacterium_sp_434_HC2	Species	235	221
gi|335053207|ref|NZ_AFIL01000016.1|:c75436-75296	Propionibacterium_sp_434_HC2	Species	141	139
gi|335054139|ref|NZ_AFIL01000041.1|:c77880-77749	Propionibacterium_sp_434_HC2	Species	132	80
gi|335054309|ref|NZ_AFIL01000044.1|:65842-65994	Propionibacterium_sp_434_HC2	Species	153	90
gi|335054434|ref|NZ_AFIL01000047.1|:12103-12642	Propionibacterium_sp_434_HC2	Species	540	493
gi|335054520|ref|NZ_AFIL01000051.1|:c25042-24929	Propionibacterium_sp_434_HC2	Species	114	73
gi|335055158|ref|NZ_AFIL01000073.1|:155425-155610	Propionibacterium_sp_434_HC2	Species	186	121
gi|355707189|ref|NZ_JH376566.1|:236054-236590	Propionibacterium_sp_5_U_42AFAA	Species	537	518
gi|355708280|ref|NZ_JH376568.1|:c63099-62986	Propionibacterium_sp_5_U_42AFAA	Species	114	304
gi|514979630|ref|NZ_KE340299.1|:c1519736-1517826	Propionibacterium_sp_HGH0353	Species	1911	702
gi|550735774|gb|AXMM01000002.1|:509428-510885	Propionibacterium_sp_KPL1844	Species	1458	700
gi|550735774|gb|AXMM01000002.1|:740-1753	Propionibacterium_sp_KPL1844	Species	1014	698

gi|550737965|gb|AXMM01000001.1|:c339754-339413	Propionibacterium_sp_KPL1844	Species	342	280
gi|552896371|ref|NZ_AXMI01000002.1|:767403-767774	Propionibacterium_sp_KPL1854	Species	372	338
gi|552896688|ref|NZ_AXMI01000003.1|:118812-119252	Propionibacterium_sp_KPL1854	Species	441	420
gi|552897324|ref|NZ_AXMI01000005.1|:788-1104	Propionibacterium_sp_KPL1854	Species	317	214
gi|552897361|ref|NZ_AXMI01000006.1|:1-107	Propionibacterium_sp_KPL1854	Species	107	72
gi|552879811|ref|NZ_AXME01000001.1|:1431752-1431913	Propionibacterium_sp_KPL2008	Species	162	83
gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	Propionibacterium_sp_KPL2008	Species	207	132
gi|552879811|ref|NZ_AXME01000001.1|:865400-865597	Propionibacterium_sp_KPL2008	Species	198	173
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	Propionibacterium_sp_KPL2008	Species	270	237
gi|552879811|ref|NZ_AXME01000001.1|:c1820429-1820292	Propionibacterium_sp_KPL2008	Species	138	97
gi|552879811|ref|NZ_AXME01000001.1|:c31864-31571	Propionibacterium_sp_KPL2008	Species	294	264
gi|552879811|ref|NZ_AXME01000001.1|:c590861-590655	Propionibacterium_sp_KPL2008	Species	207	186
gi|422323853|ref|NZ_JH370351.1|:549841-550080	Rothia	Genus	240	221

Table S2. Performance of the hidSkinPlex at ≥ 70x read depth

	BacMix		*P. acnes*		*P. granulosum*		*R. dentocariosa*	
PCR Conditions	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**
57°C, A	0.9154	0.8286	0.9634	0.7397	0.6667	0.9697	1.0000	0.9905
57°C, AQ	0.8454	0.8986	0.9662	0.8026	0.3333	0.9750	1.0000	1.0000
57°C, B	0.9712	0.8060	0.9816	0.7536	0.3333	0.9717	1.0000	1.0000
59°C, A	0.9759	0.7639	0.9882	0.7606	0.5000	0.9643	1.0000	0.9545
59°C, AQ	0.9679	0.8000	0.9821	0.7714	0.7500	0.9703	1.0000	1.0000
59°C, B	0.9784	0.8125	0.9876	0.7761	0.4000	0.9730	1.0000	0.9901

A = 8.75 nM primer concentration; B = 4.375 nM primer concentration; Q = addition of Q solution

BacMix= 1:1:1 mixture of *P. acnes*, *P. granulosum*, and *R. dentocariosa*

SN = Sensitivity; SP = Specificity

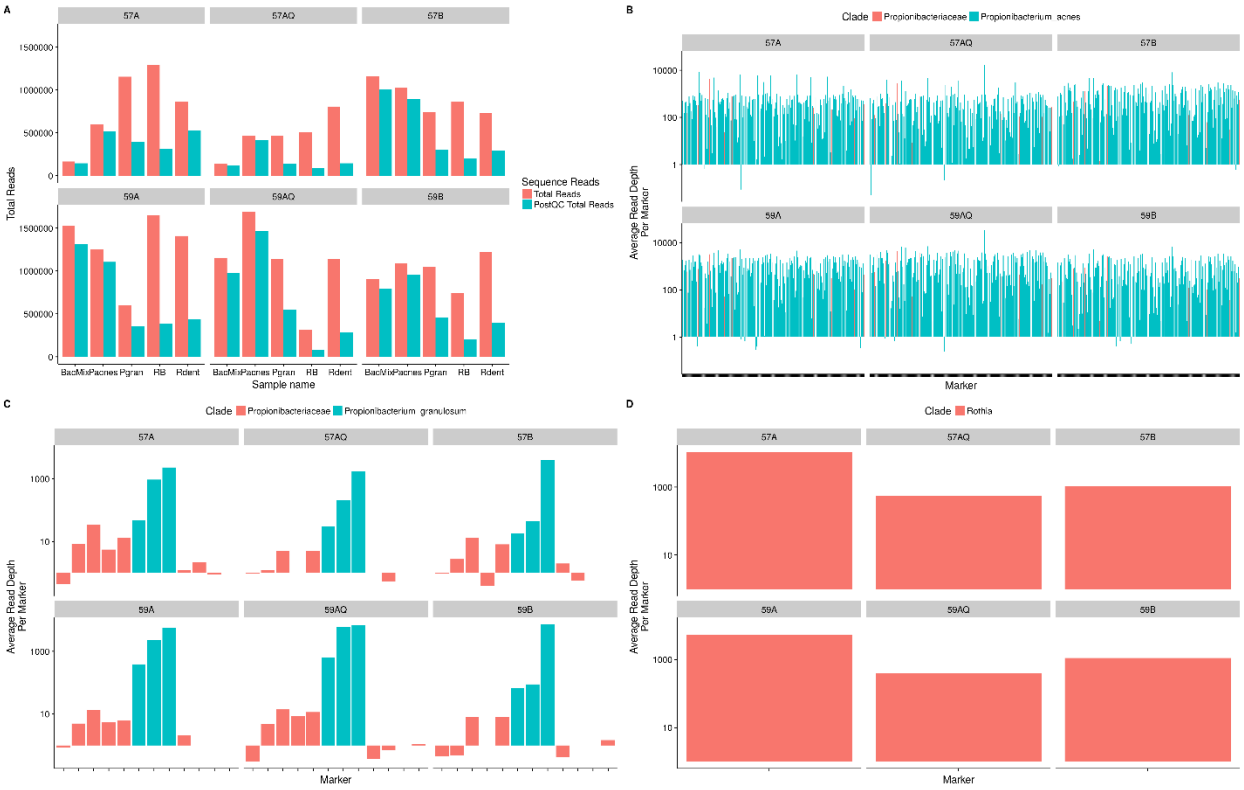

Figure S1. Performance of the hidSkinPlex assay. A) Total sequence reads, pre- and post-quality filtering, per sample for each PCR parameter. Sample names: BacMix = synthetic bacterial mixture containing equal amounts of genomic DNA from *Propionibacterium acnes*, *Propionibacterium granulosum*, and *Rothia dentocariosa*; Pacnes = *P. acnes*; Pgran = *P. granulosum*; RB = reagent blank; Rdent = *R. dentocariosa*. B) Marker read depth at each expected marker (i.e., "true positive", n = 196) for *P. acnes* on a log scale. C) Marker read depth at each expected marker (i.e., "true positive", n = 12) for *P. granulosum* on a log scale. D) Marker read depth at each expected marker (i.e., "true positive", n = 1) for *R. dentocariosa* on a log scale. PCR parameters tested, include: 57°C and 59°C annealing temperatures; A = 8.75 nM final primer concentration; B = 4.375 nM final primer concentration; Q = addition of Q solution.

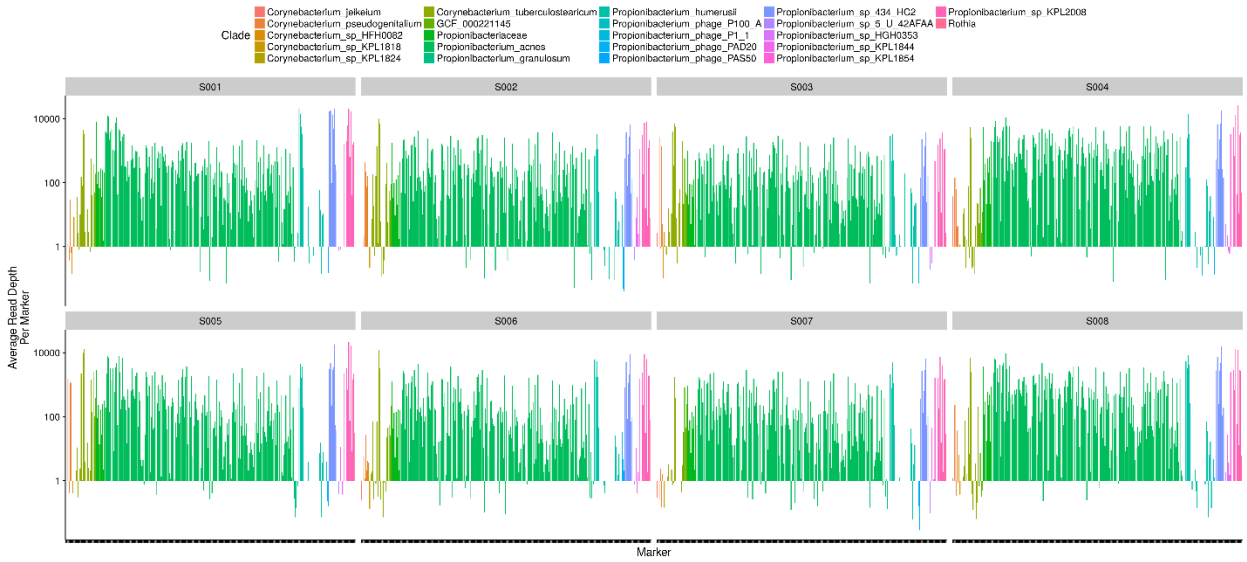

Figure S2. The average read depth at each hidSkinPlex marker present in eight individuals from the toe web/ball of the foot (Fb). Markers are ordered by clade then amplicon size on a log scale.

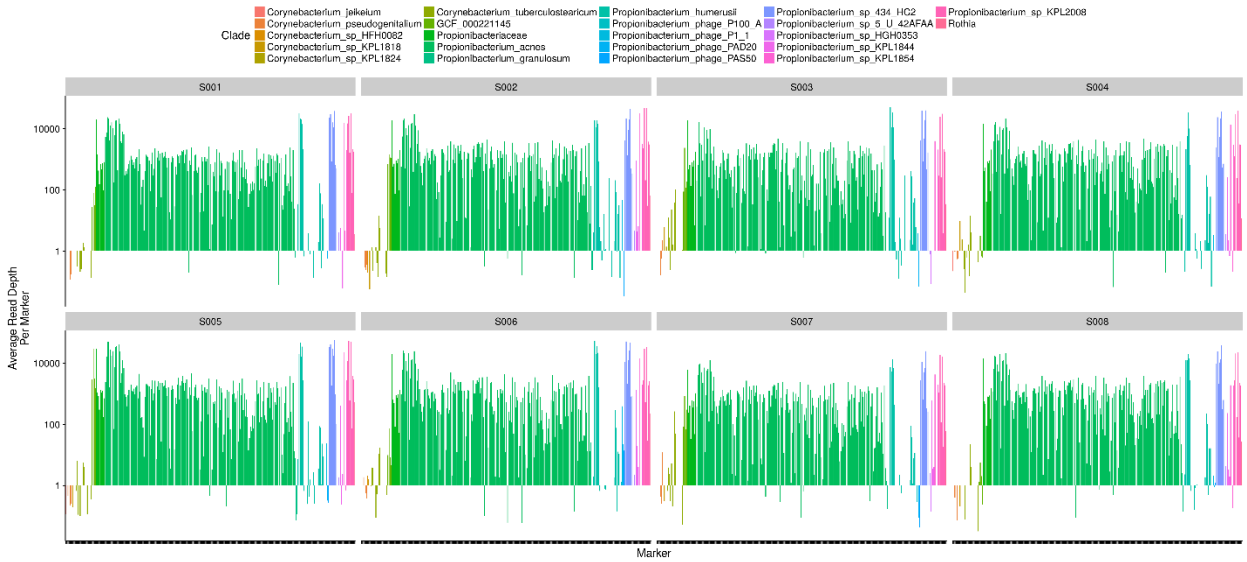

Figure S3. The average read depth at each hidSkinPlex marker present in eight individuals from the palm of the non-dominant hand (Hp). Markers are ordered by clade then amplicon size on a log scale.

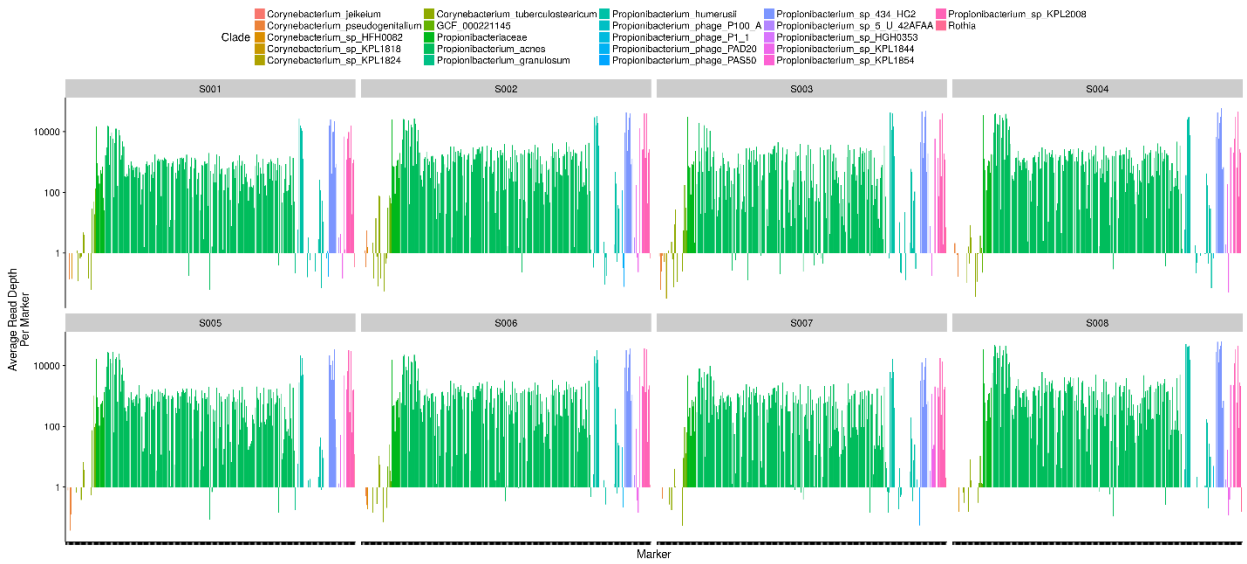

Figure S4. The average read depth at each hidSkinPlex marker present in eight individuals from the manubrium (Mb). Markers are ordered by clade then amplicon size on a log scale.

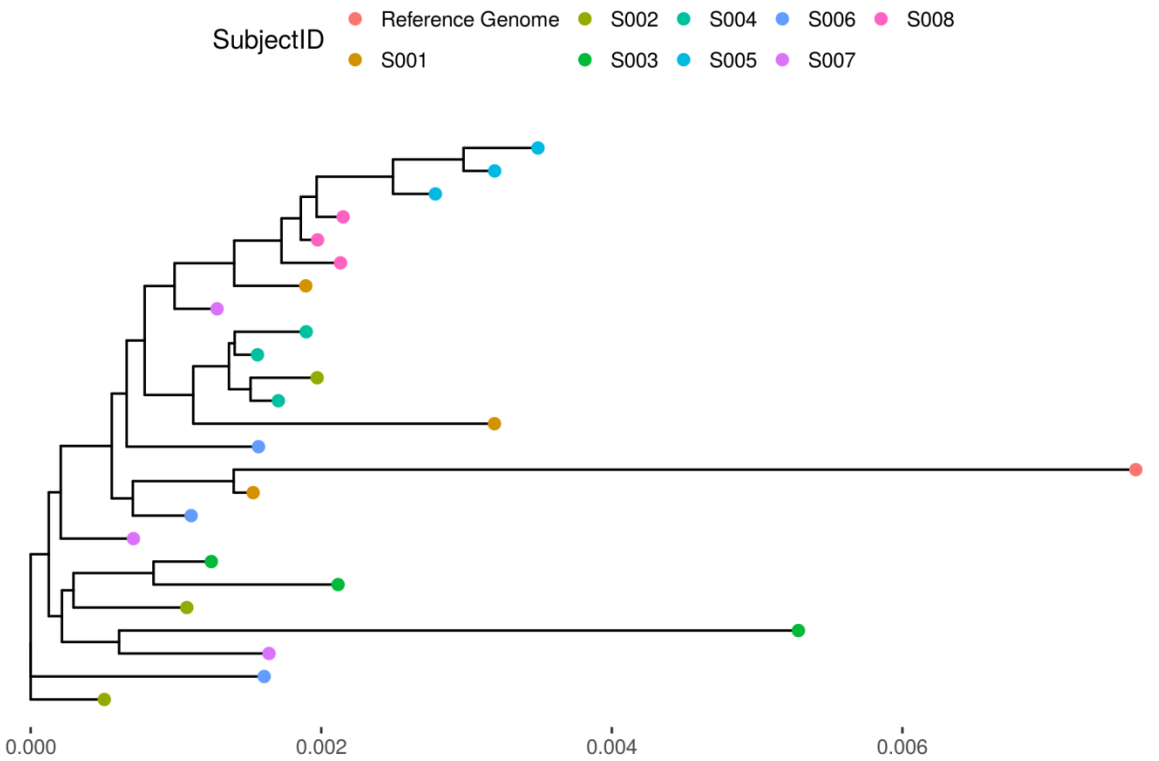

Figure S5. Maximum likelihood phylogeny of *Propionibacterium acnes* strains present on the toe web/ball of the foot (Fb) from eight individuals. The *P. acnes* phylogeny was constructed using all *P. acnes*-specific markers in the hidSkinPlex panel (n = 187).

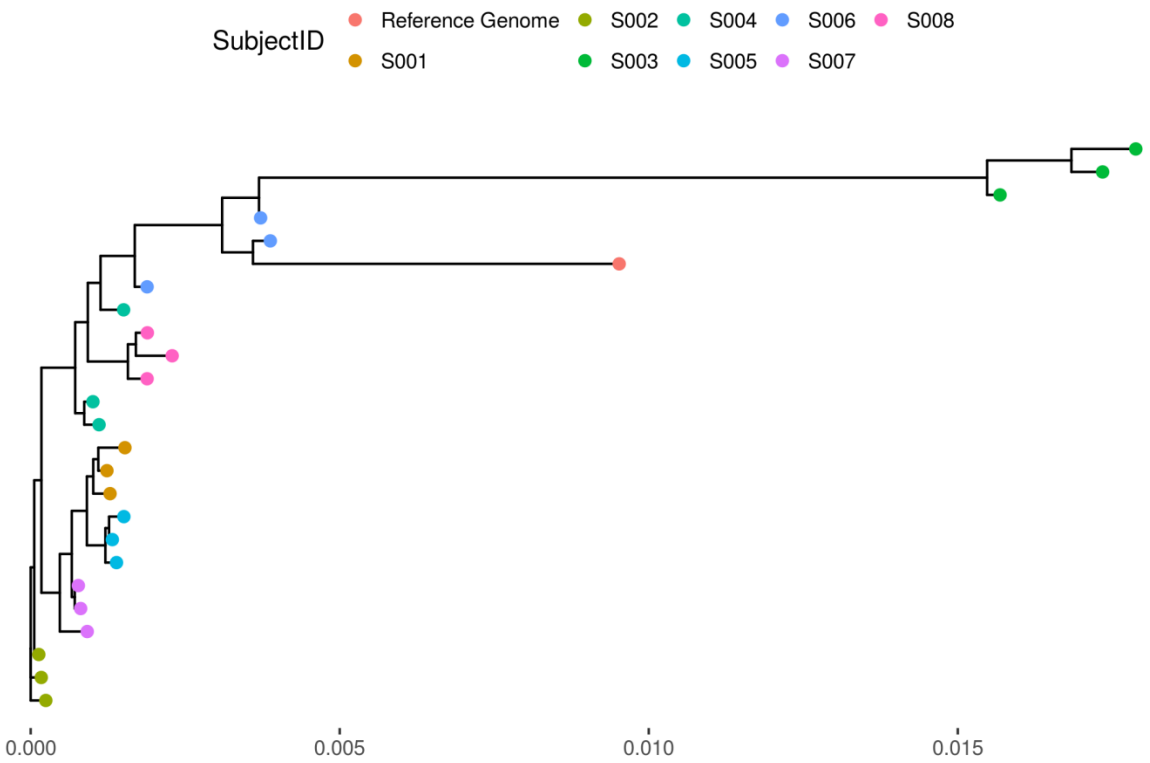

Figure S6. Maximum likelihood phylogeny of *Propionibacterium acnes* strains present on the palm of the non-dominant hand (Hp) from eight individuals. The *P. acnes* phylogeny was constructed using all *P. acnes*-specific markers in the hidSkinPlex panel (n = 187).

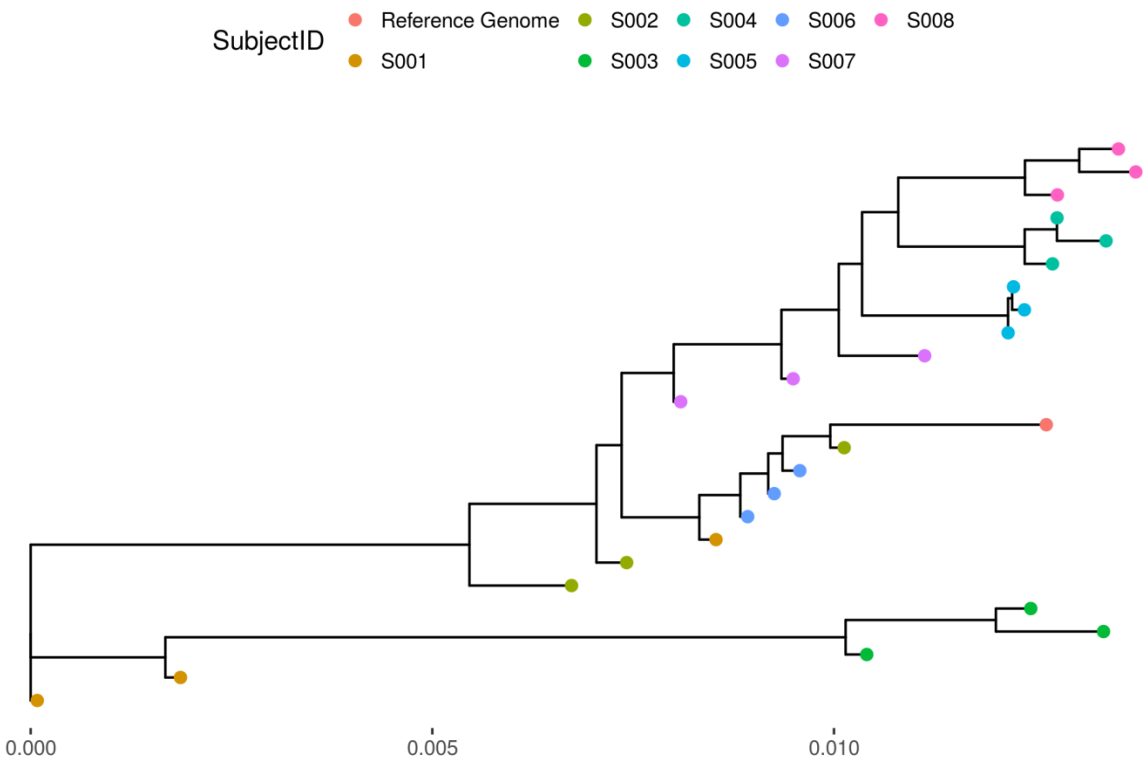

Figure S7. Maximum likelihood phylogeny of *Propionibacterium acnes* strains present on the manubrium (Mb) from eight individuals. The *P. acnes* phylogeny was constructed using all *P. acnes*-specific markers in the hidSkinPlex panel (n = 187).

Table S3. Classification Accuracies using Universal Markers

Body site Symbol	Threshold	No. of Samples	No. of Individuals	No. of Markers	No. of AttSelect Markers	% Accuracy by Random Chance	RMLR			1NN			RMLR w/AttSelect			1NN w/AttSelect		
							% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI
all	2	72	8	183	16	2.82	91.67	85.28	98.05	87.50	79.86	95.14	88.89	81.63	96.15	88.89	81.63	96.15
all	10	72	8	138	20	2.82	91.67	85.28	98.05	97.22	93.43	101.02	79.17	69.79	88.55	93.06	87.18	98.93
all	25	72	8	103	17	2.82	88.89	81.63	96.15	90.28	83.43	97.12	83.33	74.73	91.94	90.28	83.43	97.12
all	50	72	8	75	13	2.82	84.72	76.41	93.03	91.67	85.28	98.05	86.11	78.12	94.10	90.28	83.43	97.12
all	100	72	8	51	11	2.82	88.89	81.63	96.15	93.06	87.18	98.93	70.83	60.33	81.33	79.17	69.79	88.55
all	150	72	8	34	11	2.82	84.72	76.41	93.03	94.44	89.15	99.74	75.00	65.00	85.00	87.50	79.86	95.14
all	200	72	8	17	8	2.82	68.06	57.29	78.83	73.61	63.43	83.79	62.50	51.32	73.68	75.00	65.00	85.00
Fb	2	24	8	188	31	8.70	54.17	34.23	74.10	75.00	57.68	92.32	75.00	57.68	92.32	70.83	52.65	89.02
Fb	10	24	8	143	27	8.70	70.83	52.65	89.02	79.17	62.92	95.41	83.33	68.42	98.24	87.50	74.27	100.73
Fb	25	24	8	108	27	8.70	70.83	52.65	89.02	70.83	52.65	89.02	79.17	62.92	95.41	75.00	57.68	92.32
Fb	50	24	8	79	19	8.70	79.17	62.92	95.41	75.00	57.68	92.32	83.33	68.42	98.24	75.00	57.68	92.32
Fb	100	24	8	57	19	8.70	75.00	57.68	92.32	75.00	57.68	92.32	75.00	57.68	92.32	70.83	52.65	89.02
Fb	150	24	8	42	15	8.70	62.50	43.13	81.87	83.33	68.42	98.24	70.83	52.65	89.02	66.67	47.81	85.53
Fb	200	24	8	37	16	8.70	70.83	52.65	89.02	83.33	68.42	98.24	70.83	52.65	89.02	75.00	57.68	92.32
Hp	2	24	8	207	64	8.70	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00	95.83	87.84	103.83
Hp	10	24	8	188	61	8.70	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00	95.83	87.84	103.83
Hp	25	24	8	172	54	8.70	100.00	100.00	100.00	95.83	87.84	103.83	95.83	87.84	103.83	95.83	87.84	103.83
Hp	50	24	8	152	52	8.70	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00	95.83	87.84	103.83
Hp	100	24	8	134	51	8.70	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00	95.83	87.84	103.83
Hp	150	24	8	115	42	8.70	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00	95.83	87.84	103.83
Hp	200	24	8	98	38	8.70	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00	100.00	100.00	100.00
Mb	2	24	8	202	43	8.70	83.33	68.42	98.24	87.50	74.27	100.73	91.67	80.61	102.72	91.67	80.61	102.72
Mb	10	24	8	161	41	8.70	75.00	57.68	92.32	91.67	80.61	102.72	87.50	74.27	100.73	91.67	80.61	102.72
Mb	25	24	8	136	29	8.70	83.33	68.42	98.24	95.83	87.84	103.83	83.33	68.42	98.24	95.83	87.84	103.83
Mb	50	24	8	122	29	8.70	83.33	68.42	98.24	83.33	68.42	98.24	75.00	57.68	92.32	95.83	87.84	103.83
Mb	100	24	8	86	30	8.70	87.50	74.27	100.73	87.50	74.27	100.73	83.33	68.42	98.24	87.50	74.27	100.73
Mb	150	24	8	56	21	8.70	91.67	80.61	102.72	95.83	87.84	103.83	95.83	87.84	103.83	95.83	87.84	103.83
Mb	200	24	8	29	13	8.70	70.83	52.65	89.02	91.67	80.61	102.72	79.17	62.92	95.41	83.33	68.42	98.24

AttSelect = Attribute Selection

Table S4. Classification Accuracies using Non-universal Markers

Body site Symbol	Threshold	No. of Samples	No. of Individuals	No. of Markers	No. of AttSelect Markers	% Accuracy by Random Chance	RMLR			1NN			RMLR w/AttSelect			1NN w/AttSelect		
							% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI
all	2	72	8	275	29	2.82	81.94	73.06	90.83	88.89	81.63	96.15	83.33	74.73	91.94	83.33	74.73	91.94
all	10	72	8	261	25	2.82	94.44	89.15	99.74	93.06	87.18	98.93	84.72	76.41	93.03	80.56	71.41	89.70
all	25	72	8	258	31	2.82	84.72	76.41	93.03	84.72	76.41	93.03	86.11	78.12	94.10	86.11	78.12	94.10
all	50	72	8	251	22	2.82	90.28	83.43	97.12	77.78	68.17	87.38	81.94	73.06	90.83	86.11	78.12	94.10
all	100	72	8	244	27	2.82	83.33	74.73	91.94	76.39	66.58	86.20	75.00	65.00	85.00	84.72	76.41	93.03
all	150	72	8	235	25	2.82	84.72	76.41	93.03	73.61	63.43	83.79	75.00	65.00	85.00	79.17	69.79	88.55
all	200	72	8	232	22	2.82	86.11	78.12	94.10	72.22	61.88	82.57	83.33	74.73	91.94	81.94	73.06	90.83
Fb	2	24	8	263	34	8.70	58.33	38.61	78.06	87.50	74.27	100.73	62.50	43.13	81.87	83.33	68.42	98.24
Fb	10	24	8	254	45	8.70	45.83	25.90	65.77	91.67	80.61	102.72	66.67	47.81	85.53	83.33	68.42	98.24
Fb	25	24	8	240	37	8.70	45.83	25.90	65.77	75.00	57.68	92.32	70.83	52.65	89.02	66.67	47.81	85.53
Fb	50	24	8	235	38	8.70	62.50	43.13	81.87	75.00	57.68	92.32	62.50	43.13	81.87	75.00	57.68	92.32
Fb	100	24	8	220	31	8.70	54.17	34.23	74.10	62.50	43.13	81.87	75.00	57.68	92.32	70.83	52.65	89.02
Fb	150	24	8	209	20	8.70	70.83	52.65	89.02	66.67	47.81	85.53	66.67	47.81	85.53	54.17	34.23	74.10
Fb	200	24	8	199	22	8.70	62.50	43.13	81.87	58.33	38.61	78.06	70.83	52.65	89.02	62.50	43.13	81.87
Hp	2	24	8	267	71	8.70	100.00	100.00	100.00	100.00	100.00	100.00	95.83	87.84	103.83	100.00	100.00	100.00
Hp	10	24	8	247	73	8.70	95.83	87.84	103.83	100.00	100.00	100.00	87.50	74.27	100.73	95.83	87.84	103.83
Hp	25	24	8	242	61	8.70	95.83	87.84	103.83	95.83	87.84	103.83	91.67	80.61	102.72	95.83	87.84	103.83
Hp	50	24	8	233	64	8.70	100.00	100.00	100.00	95.83	87.84	103.83	91.67	80.61	102.72	100.00	100.00	100.00
Hp	100	24	8	224	66	8.70	95.83	87.84	103.83	95.83	87.84	103.83	91.67	80.61	102.72	95.83	87.84	103.83
Hp	150	24	8	219	64	8.70	91.67	80.61	102.72	95.83	87.84	103.83	91.67	80.61	102.72	100.00	100.00	100.00
Hp	200	24	8	216	63	8.70	91.67	80.61	102.72	95.83	87.84	103.83	83.33	68.42	98.24	91.67	80.61	102.72
Mb	2	24	8	258	49	8.70	66.67	47.81	85.53	83.33	68.42	98.24	70.83	52.65	89.02	87.50	74.27	100.73
Mb	10	24	8	237	44	8.70	83.33	68.42	98.24	87.50	74.27	100.73	66.67	47.81	85.53	79.17	62.92	95.41
Mb	25	24	8	232	46	8.70	75.00	57.68	92.32	83.33	68.42	98.24	75.00	57.68	92.32	83.33	68.42	98.24
Mb	50	24	8	224	40	8.70	79.17	62.92	95.41	75.00	57.68	92.32	62.50	43.13	81.87	70.83	52.65	89.02
Mb	100	24	8	219	42	8.70	70.83	52.65	89.02	70.83	52.65	89.02	70.83	52.65	89.02	75.00	57.68	92.32
Mb	150	24	8	211	49	8.70	66.67	47.81	85.53	75.00	57.68	92.32	70.83	52.65	89.02	83.33	68.42	98.24
Mb	200	24	8	209	43	8.70	70.83	52.65	89.02	70.83	52.65	89.02	75.00	57.68	92.32	70.83	52.65	89.02

AttSelect = Attribute Selection

Table S5. Classification Accuracies using Non-universal Markers for Body Site Classification

Body site Symbol	Threshold	No. of Samples	No. of Individuals	No. of Markers	No. of AttSelect Markers	RMLR			1NN			RMLR w/AttSelect			1NN w/AttSelect		
						% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI	% Accuracy	Lower 95% CI	Upper 95% CI
all	2	72	24	275	18	79.17	69.79	88.55	77.78	68.17	87.38	80.56	71.41	89.70	77.78	68.17	87.38
all	10	72	24	261	15	83.33	74.73	91.94	83.33	74.73	91.94	80.56	71.41	89.70	81.94	73.06	90.83
all	25	72	24	258	21	86.11	78.12	94.10	80.56	71.41	89.70	75.00	65.00	85.00	80.56	71.41	89.70
all	50	72	24	251	18	79.17	69.79	88.55	77.78	68.17	87.38	69.44	58.80	80.08	73.61	63.43	83.79
all	100	72	24	244	23	69.44	58.80	80.08	77.78	68.17	87.38	84.72	76.41	93.03	84.72	76.41	93.03
all	150	72	24	235	17	79.17	69.79	88.55	73.61	63.43	83.79	76.39	66.58	86.20	83.33	74.73	91.94
all	200	72	24	232	15	77.78	68.17	87.38	73.61	63.43	83.79	77.78	68.17	87.38	84.72	76.41	93.03

AttSelect = Attribute Selection

Supplemental Files

File S1. - (Excel workbook) – Available online upon publication.

SUMMARY

Genetic Profiling of Skin Microbiomes for Forensic Human Identification

The microbial forensics field has expanded to include numerous applications beyond the traditional focus of biodefense and biocrime attribution due to the technological advancements in massively parallel sequencing and bioinformatics, leading to increased throughput and speed at which microbial genomes and metagenomes are sequenced. The increase in bacterial, archaeal, viral, and microbial eukaryotic genomes in public databases has spurred the advent and growth of new fields, such as metagenomics, comparative genomics, and microbial forensics. Metagenomics and comparative genomic capabilities have fueled the expansion of the microbial forensics field from a strictly bioterrorism/biocrime focus to include human identification, post-mortem interval, trace microbial evidence, and the potential for recent geolocation. The primary goal of this dissertation was to develop a novel metagenomics sequencing method to profile skin microbiomes for forensic human identification.

In **Chapter 1** a preliminary study presents a novel tool, AutoCurE, for maintaining and curating a local bacterial genome database. This study was undertaken to overcome several inconsistencies and errors in genome data and metadata, initially observed, when constructing a local bacterial genome database for use for downstream comparative genomic studies. In this study, all publically-available complete bacterial genomes (n=2,769) were downloaded from the NCBI ftp site, along with current genome reports and associated metadata for each downloaded genome. After three rounds of manual curation 189 genomes were removed from the database due to several inconsistencies and errors, including genomes identified as archaea, draft sequences, and only plasmid sequences present. Additional errors identified included genomes not present in genomes reports (or present in reports but not available for download), inconsistencies between genus and species names, missing or discontinued accession numbers, change from complete status, missing complete reference assemblies, and erroneous sequence files included in the

genome folder. The identification of errors is imperative to prevent or reduce incorrect characterizations of sequence data.

One-by-one manual curation of the data is time consuming and tedious and errors may still be missed. An automated curation tool in Excel (AutoCurE) was developed and validated for curation of local bacterial genome databases. AutoCurE includes two Excel workbooks, the AutoCurE Genome Filename Tool and the AutoCurE Genome Report Tool which generate flags for nine categories related to accession numbers, BioProject/UID, genus and species consistency, archaea, sequence files present, and draft or partial sequences. The main features of AutoCurE include print list directory of downloaded genomes and file paths, retrieve RefSeq accession and sequence file description, parse metadata from genome reports and downloaded sequence files, and file manipulation to eliminate manual searching within directories. AutoCurE provides an easy-to-use tool for non-programmers to curate local bacterial genome databases.

In the second section of this dissertation, Chapters 2 and 3 present the studies conducted to test the hypothesis that genes from stable, universal microbial species from the core skin microbiome can differentiate skin microbiomes of individuals and be applied towards forensic human identification purposes.

Chapter 2, presented a novel approach for characterization of skin microbiomes to identify individual-specific signatures that can be used for human identification. A publically-available shotgun metagenomic sequence dataset, comprised of data generated from spatially and temporally sampled skin microbiomes, was mined to identify stable features which could be used to differentiate individuals. Skin microbiome samples in the dataset were collected from 14 body sites from 12 healthy individuals, sampled at three time points over a period of ~ 3years. Two skin microbiome feature types (i.e., variables used for classification), *Propionibacterium acnes*

pangenome gene presence/absence and nucleotide diversity of clade-specific markers, were assessed with regularized multinomial logistic regression (RMLR) and 1-nearest-neighbor classification (1NN) to compare the accuracy of each feature type for classification of skin microbiomes to their host individuals. Conditional binomial logistic regression was used to model the log odds of a correct classification as a linear function to compare which factors (i.e., body site, feature type, and classification method) may influence the probability of a correct classification. Nucleotide diversity of clade-specific markers contributed significantly greater to accuracy, by an estimated 28%, than classification using *P. acnes* presence/absence features. Accuracies were as high as 100% for samples from the cheek, inguinal crease, and popliteal fossa. Body sites with likely greater forensic relevance, the manubrium (shirt collar) and the hand (palm), yielded high classification accuracies (97% and 96% accuracy, respectively, using 1NN).

Attribute selection also was performed with RMLR and 1NN to select for reduced subsets of features which have similar predictive power as using all markers. The subset of attribute selected markers (i.e., reduced marker sets) performed similar to using full feature sets and thus did not compromise classification accuracy. As such, feature selection was used to identify candidate markers that could constitute a multiplex targeted sequencing panel. Moreover, attribute selection was performed independent of classifier type, and as such identified features potentially informative for other supervised learning algorithms which may be assessed in future studies. The better performance of nucleotide diversity features than presence/absence features and use of reduced subsets of markers without effect on classification accuracy support the potential to evaluate targeted enrichment approaches for skin microbiome profiling. Shotgun metagenomic data evaluated in this study demonstrated stochastic effects when analyzing body sites such as the foot, a body site removed from this *in silico* study due to low abundant and highly variable markers

across individuals. Targeted enrichment sequencing methods, such as targeting clade-specific markers, may provide more uniform coverage of informative sites for improved classification capabilities of skin microbiomes to use for forensic identification purposes.

Finally, **Chapter 3** described the development and evaluation of a novel targeted metagenomic sequencing method to generate individual-specific skin microbiome profiles to use for human identification. Clade-specific markers, selected by attribute selection, described in Chapter 2, were developed into a multiplex amplification assay and integrated into library preparation to produce a targeted sequencing method, the hidSkinPlex, for skin microbiome profiling. The hidSkinPlex is comprised of 282 bacterial (and 4 phage) markers from 22 family-, genus-, species- and subspecies-level clades. The hidSkinPlex initially was evaluated using purified nucleic acids from three bacterial control samples, *Propionibacterium acnes*, *Propionibacterium granulosum*, and *Rothia dentocariosa*, which are targets represented in the panel. The multiplex was evaluated for optimal annealing temperature, primer concentration, and addition of Q-solution (Qiagen). The performance of the hidSkinPlex was assessed by calculating the sensitivity and specificity of the markers and uniformity of read depth across panel markers. The hidSkinPlex was further evaluated for predictive power by assessing the performance of classification algorithms, RMLR and 1NN, using nucleotide diversity of hidSkinPlex markers enriched from skin microbiome samples collected from eight individuals. Skin swabs were collected from eight individuals and three body sites (i.e., foot (Fb), hand (Hp), and manubrium (Mb)) in triplicate. RMLR and 1NN were performed to predict which skin microbiome sample was collected from the correct individual host. Skin microbiomes could be correctly attributed to their respective donors with up to 92% (Fb), 96% (Mb), and 100% (Hp) accuracy. Samples were classified with up to 97% accuracy when the body site was unknown, suggesting hidSkinPlex

markers may be informative across multiple body sites and useful in forensic settings when the body site origin may be unknown. Additionally, body site origin could be predicted with up to 86% accuracy.

Finally, a case study was conducted to highlight the potential to use microbiome profiles independently or in conjunction with human profiles for low-biomass samples. Human STR and SNP profiles were generated from skin swabs from one study subject to evaluate the percentage of alleles which could be detected from the skin swabs. Three samples from one foot replicate and two replicates from the manubrium yielded full or nearly full (92-100%) profiles compared to a reference buccal sample. All replicates sampled from the hand yielded partial profiles with only 32-52% alleles detected, the same samples which produced up to 100% classification accuracy using full hidSkinPlex profiles. This case study demonstrates the potential to use microbiome profiles for human identification independent of or in conjunction with partial human profiles from skin contact or touch evidentiary samples, sample types likely to yield even fewer human alleles and potentially more comprehensive microbial profiles.

The studies described in this dissertation contribute novel findings and tools to the expanding field of microbial forensics and metagenomics. Chapter 1 introduced a tool for the non-programmer to use to identify errors in NCBI data and allow for generating a curated local bacterial genome database. The tool and its use highlighted the types of inconsistencies and errors which may be present in public genome databases. Chapter 2 presented a novel approach to characterize individual-specific strain-level signatures and two feature types from skin microbiomes to use with supervised learning to attribute microbiome samples to their individual hosts. Additionally, the study described in Chapter 2 identified a set of candidate markers for potential development of a skin microbiome forensic human identification panel. Lastly, Chapter 3 described the development

of the hidSkinPlex, a novel targeted sequencing panel for forensic human identification using skin microbiome profiles. Future studies should focus on further optimization of the hidSkinPlex and generation of population studies to further assess the stability and diversity of the skin microbiome as applicable to forensic human identification and microbial trace evidence.

CONCLUSIONS AND FUTURE DIRECTIONS

Genetic Profiling of Skin Microbiomes for Forensic Human Identification

Advancements in massively parallel sequencing and bioinformatics have opened the door to newer applications in the forensic sciences. Microbial forensics, previously with the sole focus of biodefense and biocrime attribution, has expanded to provide more tools for traditional forensic investigations (1). Newer microbial forensic applications have focused on post-mortem interval (i.e., using microbial signatures from human decomposition to predict the interval since time-of-death) (2, 3), infection source tracking (e.g., determining patient zero in cases of deliberate or negligent transmission of HCV and HIV) (4–6), forensic identification of trace soil evidence (7), body fluid identification (8, 9), identifying and tracking drug users and networks (10), and human identification (11–13). Trace evidence now includes microbial signatures from persons, surfaces or even the air, relying more so on culture-independent methods (i.e., metagenomics) to attribute trace microbial evidence to a perpetrator(s) or other source or origin. The work presented in this dissertation directly contributes to the microbial forensics toolbox, specifically developing novel targeted sequencing and bioinformatics methods to generate and classify skin microbiome signatures for forensic human identification applications.

Public genome databases have substantially increased as sequencing technologies have advanced producing higher throughout instruments at lower costs. As such, consortiums seek to increase of the number of prokaryotic genomes, both in number and phylogenetic diversity that will be publicly accessible (14–17). At the time AutoCurE was developed, during 2014-2015, there were > 2,700 complete bacterial and archaeal genomes publically available. Two years later and at the time of this writing, there are > 8,100 complete prokaryotic genomes in the NCBI Genome database (18), which is a substantial increase in a 2-3 year period. Shortly after AutoCurE was published, NCBI made vast improvements to its ftp site and genome database organization, resolving many issues identified by Schmedes et al. (19). The most notable changes were the

addition of links, to GenBank (20) and RefSeq (21) assemblies, included directly in the genome report and the physical location and organization of GenBank (20) and RefSef (21) sequence files. These changes dramatically reduce the need for manual curation of local databases, although genome files downloaded from public databases should always undergo some level of quality control prior to use. Additionally, as the number of sequenced genomes has substantially increased, and will likely continue to do so, a tool such as AutoCure may no longer be the most appropriate choice due to the memory and computational constraints of Excel.

Shifts in the field of microbial genomics, especially involving "big data", are creating a demand for bioinformaticians in the typical molecular biology laboratory. For laboratories which lack the funding or need for full-time bioinformaticians, commercial bioinformatics software platforms are more readily available for non-programmers and novice bioinformaticians. While AutoCurE can still be used to maintain and curate a local bacterial database, the need for more advanced solutions for database storage and manipulation are more readily apparent than they were just a few years ago. Sequence data files, especially shotgun metagenomic files, are getting larger as sequencing platforms increase throughput. Resources such as the NCBI SRA (22) are imperative for sequence storage and back up and resource sharing among the scientific community. Continued support and guidance from standards consortiums to standardize metadata and database quality control are necessary to ensure standardization of sequence metadata with the expansion of genome databases.

The primary studies of this dissertation, in Chapters 2 and 3, focused on developing a method to utilize skin microbiome signatures for forensic identification, a method only recently possible to develop, due to the advancements in metagenomics and bioinformatics and with access to publically-available data (i.e., on the NCBI SRA (22)). Schmedes et al. (23) presented the first

targeted panel, developed into the hidSkinPlex, designed specifically to generate individual skin microbiome genetic profiles to use for forensic human identification. The hidSkinPlex panel and multiplex assay were designed as an alternative to metagenomic sequencing methods that traditionally targeted 16S rRNA or shotgun metagenomic sequencing. The hidSkinPlex improves upon the limitations of 16S rRNA and shotgun sequencing by capturing only informative sites down to strain-level resolution, which can maximize coverage and read depth, thus reducing stochastic effects. The use of skin microbiome prediction using the hidSkinPlex within a supervised learning context demonstrated the capability to predict individual identification using skin microbiome profiles.

In Chapters 2 and 3, the limitations of using shotgun metagenomic data for identification are highlighted and demonstrate enriched targets can provide more uniform coverage of universal markers across body sites. Targeted enrichment provides the capability to identify individuals using samples from a body site with reduced stochastic effects (i.e., the foot), and to identify individuals using samples across the body, regardless of body site as well as predict body site origin. One of the more surprising and substantial findings from these studies was the ability to attribute skin microbiome samples to their respective host with up to 100% accuracy using samples collected from the hand. The hand is one of the most forensically relevant sites, regarding “touch DNA” samples, and as such this finding is significant for the potential use of skin microbiome profiling independent of or in conjunction with traditional human forensic profiles to assist in criminal investigations, such as robberies, homicides, and sexual assaults.

The results from these studies provide a preliminary proof-of-concept and a multiplex panel to assess the stability and diversity of the skin microbiome and how it relates to human identification. Future studies should evaluate the hidSkinPlex on a larger number of individuals

and skin body sites, sampled over time. The hidSkinPlex will next be evaluated on group of 50 individuals, assessing the ability to differentiate subjects using samples collected from the foot, hand, and manubrium prior to expanding testing to larger population studies. After panel optimization, population data should be generated from at least 200 samples each from major US population groups, including Caucasian, African-American, Hispanic, Native American, and Asian, and population groups from different geographical locations.

Future studies also should focus on optimization of the hidSkinPlex panel, including marker reduction/inclusion and primer redesign. The hidSkinPlex markers were selected from multiple body sites, from two different sample sets using various threshold parameters and sample inclusion criteria. This marker selection process allowed for redundancy of marker diversity and inclusion to assess the performance of each marker in the case certain markers failed to amplify and to determine which markers may be redundant or less informative. Future development of the panel should focus on removing markers which are less informative and/or contribute noise to the system, the addition of markers to improve classification accuracy for the foot and body site origin, and to identify the specific regions within each marker which provide the most individualizing information. Therefore, primers should be redesigned to capture these more informative regions, likely reducing the amplicon size and creating more uniform amplicon sizes which may be able capture entire markers within a single sequence read.

The hidSkinPlex also should be assessed on samples from additional skin body sites, including a subset of the 17 body sites from Oh et al. (24). Markers included in the hidSkinPlex panel were selected from 14/17 body sites (from Oh et al. (24)), and therefore the performance of the panel could be evaluated to determine if classification accuracies are comparable or better than the Fb, Hp, and Mb assessed in this study. Additional markers also should be evaluated for

inclusion in the hidSkinPlex for body site identification capabilities. In Chapter 3, body site origin could be predicted with up to 86% accuracy. Additional markers selected using methods similar to those described in Chapter 2, from shotgun data, may provide additional resolution of the hidSkinPlex for body site identification. Body site prediction using the hidSkinPlex would provide an additional tool for forensic investigations, as the body site of origin likely may be unknown for some evidentiary samples. Body site origin could help corroborate testimonies or produce investigative leads depending on the nature of the investigation and case.

Additional analysis methods and supervised learning algorithms also should be assessed to develop a standardized bioinformatics pipeline and interpretation guidelines for using the hidSkinPlex in the forensic setting. In this study, RMLR and 1NN were evaluated for assessing predictive power using both shotgun metagenomic and targeted enriched data. Additional algorithms such as support vector machines, logistic regression using the lasso parameter, K-nearest-neighbor ($K > 1$), and random forest classification should be assessed once larger sample sizes are generated. Also, other features from the data should be compared to nucleotide diversity to identify the most useful feature-type for human identification. SNPs or haplotype generation may provide more discrimination than nucleotide diversity, by identifying genetic diversity at specific loci/base positions within a marker. While nucleotide diversity provides measures of marker heterozygosity at the strain-level, SNP profiling and haplotypes assessed using phylogenetic methods may provide greater accuracy for evaluating microbial signatures for human identification. There is substantially more information within the sequence of the features than is captured by nucleotide diversity. Indeed, this area of SNP profiling should be the next focus to bring human identification by microbiome analyses to fruition.

Finally, future studies should address the major influences on skin microbiome composition, both genetic and environmental, including influence from frequent contact with cohabitating individuals in the same household, such as spouses, family members, and even pets. Previous studies have shown that microbiome composition may be heavily influenced by cohabitating individuals (25, 26). The hidSkinPlex could be used to evaluate the effects cohabitating couples and family members have on the skin microbiome composition and diversity and assess the capability to distinguish skin microbiomes from individuals within the same household. In Chapter 3, trace human alleles on the skin of subject S001 were detected, likely from a male donor with frequent contact with this subject. Further studies also will evaluate the level of trace human DNA on cohabitating family members and assess how long these trace profiles exist on a person after contact with another individual. Both human and microbial trace DNA may serve as evidence in cases of recent geolocation or contact and sexual assault investigations.

In this dissertation, a novel targeted metagenomics sequencing method, the hidSkinPlex, is presented as a new tool for forensic human identification using skin microbiomes. Future studies should focus on the optimization of the hidSkinPlex, further expanding the capabilities of identification and body site prediction of microbial signatures transferred to touched evidentiary items, vastly expanding current, limited forensic testing capabilities using touch DNA. As sequencing and bioinformatics technologies continue to advance, additional tools and methodologies will contribute to the expansion of microbial forensic capabilities and be used in the standard forensic workflow for both criminal and civil investigations.

REFERENCES

1. Schmedes SE, Sajantila A, Budowle B. 2016. Expansion of Microbial Forensics. J Clin Microbiol 54:1964–1974.

2. Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S, Tomberlin JK. 2014. The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. Int J Legal Med 128:193–205.

3. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale J V., DeBruyn JM, Lents NH, Oh J, Byrd AL, Park M, Kong HH, Segre JA, Bashan A, Gibson T, Friedman J, Carey V, Weiss S, Hohmann E, Jortha P, Turner K, Gumus P, Nizam N, Buduneli N, Whiteley M, Hyde E, Haarmann D, Lynne A, Bucheli S, Petrosino J, Metcalf JL, Parfrey LW, Gonzalez A, Lauber CL, Knights D, Ackermann G, Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S, Tomberlin JK, Pechal JL, Crippen TL, Tarone AM, Lewis AJ, Tomberlin JK, Benbow ME, Carter DO, Metcalf JL, Bibat A, Knight R, Cobaugh KL, Schaeffer SM, DeBruyn JM, Finley SJ, Pechal JL, Benbow ME, Robertson BK, Javan GT, Hauther KA, Cobaugh KL, Jantz LM, Sparer TE, DeBruyn JM, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Guyer M, Michaud J-P, Gaétan M, Bishop C, Hill MO, Alpaydin E, Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Basak D, Srimanta P, Patranabis DC, Hoerl AE, Kennard RW, Liaw A, Matthew W, Saeys Y, Inza I, Larrañaga P, Guyon I, Elisseeff A, Metcalf JL, Xu ZZ, Weiss S, Lax S, Treuren W Van, Hyde ER. 2016. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS One 11:e0167370.

4. Metzker ML, Mindell DP, Liu X-M, Ptak RG, Gibbs RA, Hillis DM. 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proc Natl Acad Sci U S A 99:14292–14297.

5. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. 2010. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proc Natl Acad Sci U S A 107:21242–21247.

6. González-Candelas F, Bracho MA, Wróbel B, Moya A. 2013. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. BMC Biol 11:76.

7. Santiago-Rodriguez T, Cano R. 2016. Soil Microbial Forensics. Microbiol Spectr 1–15.

8. Giampaoli S, Berti A, Valeriani F, Gianfranceschi G, Piccolella A, Buggiotti L, Rapone C, Valentini A, Ripani L, Romano Spica V. 2012. Molecular identification of vaginal fluid by microbial signature. Forensic Sci Int Genet 6:559–564.

9. Choi A, Shin K-J, Yang WI, Lee HY. 2014. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med 128:33–41.

10. Quagliarello B, Cespedes C, Miller M, Toro A, Vavagiakis P, Klein RS, Lowy FD. 2002. Strains of Staphylococcus aureus obtained from drug-use networks are closely linked. Clin Infect Dis 35:671–677.

11. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic

identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481.

12. Leake SL, Pagni M, Falquet L, Taroni F, Greub G. 2016. The salivary microbiome for differentiating individuals: proof of principle. Microbes Infect 1–7.

13. Franzosa E a., Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci 112:E2930–E2938.

14. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060.

15. Kyrpides NC, Woyke T, Eisen JA, Garrity G, Lilburn TG, Beck BJ, Whitman WB, Hugenholtz P, Klenk H-P. 2014. Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. Stand Genomic Sci 9:1278–1296.

16. Kyrpides NC, Hugenholtz P, Eisen J a, Woyke T, Göker M, Parker CT, Amann R, Beck BJ, Chain PSG, Chun J, Colwell RR, Danchin A, Dawyndt P, Dedeurwaerdere T, DeLong EF, Detter JC, De Vos P, Donohue TJ, Dong X-Z, Ehrlich DS, Fraser C, Gibbs R, Gilbert J, Gilna P, Glöckner FO, Jansson JK, Keasling JD, Knight R, Labeda D, Lapidus A, Lee J-S, Li W-J, Ma J, Markowitz V, Moore ERB, Morrison M, Meyer F, Nelson KE, Ohkuma M, Ouzounis CA, Pace N, Parkhill J, Qin N, Rossello-Mora R, Sikorski J, Smith D, Sogin M, Stevens R, Stingl U, Suzuki K-I, Taylor D, Tiedje JM, Tindall B, Wagner M, Weinstock G, Weissenbach J, White O, Wang J, Zhang L, Zhou Y-G, Field D, Whitman WB, Garrity GM, Klenk H-P. 2014. Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. PLoS Biol 12:e1001920.

17. Whitman WB, Woyke T, Klenk H-P, Zhou Y, Lilburn TG, Beck BJ, De Vos P, Vandamme P, Eisen JA, Garrity G, Hugenholtz P, Kyrpides NC. 2015. Genomic Encyclopedia of Bacterial and Archaeal Type Strains, Phase III: the genomes of soil and plant-associated and newly described type strains. Stand Genomic Sci 10:26.

18. NCBI Genome. 2017. https://www.ncbi.nlm.nih.gov/genome/browse/. Date accessed: 2017-08-19.

19. Schmedes SE, King JL, Budowle B. 2015. Correcting Inconsistencies and Errors in Bacterial Genome Metadata Using an Automated Curation Tool in Excel (AutoCurE). Front Bioeng Biotechnol 3:138.

20. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. Nucleic Acids Res 43:D30–D35.

21. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res 43:D599–D605.

22. Kodama Y, Shumway M, Leinonen R. 2012. The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56.

23. Schmedes SE, Woerner AE, Budowle B. 2017. Forensic human identification using skin microbiomes. Appl Environ Microbiol (under review).

24. Oh J, Byrd AL, Park M, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854–866.

25. Song SJ, Lauber C, Costello EK, Lozupone C a, Humphrey G, Berg-Lyons D, Caporaso JG, Knights D, Clemente JC, Nakielny S, Gordon JI, Fierer N, Knight R. 2013. Cohabiting family members share microbiota with one another and with their dogs. Elife 2:e00458.

26. Ross AA, Doxey AC, Neufeld JD. 2017. The Skin Microbiome of Cohabiting Couples. mSystems 2:e00043-17.

REFERENCES

Genetic Profiling of Skin Microbiomes for Forensic Human Identification

Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, Goedert JJ, Hayes RB, Yang L. 2013. Human Gut Microbiome and Risk of Colorectal Cancer. J Natl Cancer Inst 105:1907–1911.

Amid C, Birney E, Bower L, Cerdeño-Tárraga A, Cheng Y, Cleland I, Faruque N, Gibson R, Goodgame N, Hunter C, Jang M, Leinonen R, Liu X, Oisel A, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Rivière S, Rossello M, Senf A, Smirnov D, Ten Hoopen P, Vaughan D, Vaughan R, Zalunin V, Cochrane G. 2012. Major submissions tool developments at the European nucleotide archive. Nucleic Acids Res 40:D43-47.

Asai T, Zaporojets D, Squires C, Squires CL. 1999. An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc Natl Acad Sci U S A 96:1971–1976.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. Nucleic Acids Res 43:D30–D35.

Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG. 2015. Host genetic variation impacts microbiome composition across human body sites. Genome Biol 16:191.

Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, D. Lieber A, Wu F, Perez-Perez GI, Chen Y, Schweizer W, Zheng X, Contreras M, Dominguez-Bello MG, Blaser MJ. 2016. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med 8:1–14.

Budowle B, Schutzer SE, Einseln A, Kelley LC, Walsh AC, Smith JAL, Marrone BL, Robertson J, Campos J. 2003. Building microbial forensics as a response to bioterrorism. Science 301:1852–1853.

Budowle B, Van Daal A. 2008. Forensically relevant SNP classes. Biotechniques 44:603–610.

Budowle B, Eisenberg AJ, van Daal A. 2009. Validity of low copy number typing and applications to forensic science. Croat Med J 50:207–217.

Califf K, Gonzalez A, Knight R, Caporaso JG. 2014. The Human Microbiome : Getting Personal. Microbe 9:410–415.

Capone KA, Dowd SE, Stamatas GN, Nikolovski J. 2011. Diversity of the human skin microbiome early in life. J Invest Dermatol 131:2026–2032.

Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260–270.

Choi A, Shin K-J, Yang WI, Lee HY. 2014. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med 128:33–41.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104:19428–19433.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642.

Collins PJ, Hennessy LK, Leibelt CS, Roby RK, Reeder DJ, Foxall PA. 2004. Developmental Validation of a Single-tube Amplification of the 13 CODIS STR Loci, D2S1338, D19S433, and Amelogenin: The AmpFlSTR Identifiler PCR Amplificaiont Kit. J Forensic Sci 49:JFS2002195.

Conlan S, Mijares LA, Becker J, Blakesley RW, Bouffard GG, Brooks S, Coleman H, Gupta J, Gurson N, Park M, Schmidt B, Thomas PJ, Otto M, Kong HH, Murray PR, Segre JA. 2012. Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome Biol 13:R64.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694–1697.

Davis C, Peters D, Warshauer D, King J, Budowle B. 2015. Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing. Leg Med 17:123–127.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072.

Dorai-Raj S. 2014. binom: Binomial Confidence Intervals for Several Parameterizations. R package version 1.1-1.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

Edwards A, Civitello A, Hammond HA, Caskey CT. 1991. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet 49:746–756.

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R. 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241–253.

Ensenberger MG, Lenz KA, Matthies LK, Hadinoto GM, Schienman JE, Przech AJ, Morganti MW, Renstrom DT, Baker VM, Gawrys KM, Hoogendoorn M, Steffen CR, Martín P, Alonso A, Olson HR, Sprecher CJ, Storts DR. 2016. Developmental validation of the PowerPlex® Fusion 6C System. Forensic Sci Int Genet 21:134–144.

Federhen S, Clark K, Barrett T, Parkinson H, Ostell J, Kodama Y, et al. 2014. Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and assembly records. Stand. Genomic Sci. 9:1275–1277. doi:10.4056/sigs.4851102

Fierer N, Hamady M, Lauber CL, Knight R. 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci U S A 105:17994–17999.

Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481.

Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, Schoenfeld D, Nomicos E, Park M, NIH Intramural Sequencing Center Comparative Sequencing Program, Kong HH, Segre JA. 2013. Topographic diversity of fungal and bacterial communities in human skin. Nature 498:367–370.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512. doi:10.1126/ science.7542800

Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vázquez-Baeza Y, Gonzalez A, Knight R, Dunn RR, Fierer N. 2014. Temporal variability is a personalized feature of the human microbiome. Genome Biol 15:531.

Flores S, Sun J, King J, Budowle B. 2014. Internal validation of the GlobalFiler??? Express PCR Amplification Kit for the direct amplification of reference DNA samples on a high-throughput automated workflow. Forensic Sci Int Genet 10:33–39.

Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, Gouilh MA, Pariente K, Segondy M, Burguière A, Manuguerra JC, Caro V, Eloit M. 2012. Human skin Microbiota: High diversity of DNA viruses identified on the human skin by high throughput sequencing. PLoS One 7:e38499.

Fox GE, Wisotzkey JD, Jurtshuk, Jr. P. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170.

Frank E, Hall MA, Witten IH. 2016. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," p. . In Kauffmann, M (ed.), The WEKA WorkbenchFourth Edi.

Franzosa EA., Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci 112:E2930–E2938.

Fraser-Liggett CM. 2005. Insights on biology and evolution from microbial genome sequencing. Genome Res 15:1603–1610.

Giampaoli S, Berti A, Valeriani F, Gianfranceschi G, Piccolella A, Buggiotti L, Rapone C, Valentini A, Ripani L, Romano Spica V. 2012. Molecular identification of vaginal fluid by microbial signature. Forensic Sci Int Genet 6:559–564.

Gilbert JA, Jansson JK, Knight R. 2014. The Earth Microbiome project: successes and aspirations. BMC Biol 12:69.

Goga H. 2012. Comparison of bacterial DNA profiles of footwear insoles and soles of feet for the forensic discrimination of footwear owners. Int J Legal Med 126:815–823.

González-Candelas F, Bracho MA, Wróbel B, Moya A. 2013. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. BMC Biol 11:76.

Grice EA, Kong HH, Renaud G, Young AC, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre J a. 2008. A diversity profile of the human skin microbiota. Genome Res 18:1043–1050.

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190–1192.

Grice EA, Segre JA. 2012. The Human Microbiome: Our Second Genome. Annu Rev Genomics Hum Genet 13:151–170.

Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, Sanmiguel AJ, Minot S, Bushman FD, Grice EA, Grice A. 2015. The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. MBio 6:e01578-15.

Hares DR. 2015. Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet 17:33–34.

Holland MM, Parsons TJ. 1999. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Sci Rev. 11:21-50.

Human Microbiome Jumpstart Reference Strains Consortium. 2010. A catalog of reference genomes from the human microbiome. Science 328:994–999.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature 486:215–221.

Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L. 2010. Short-term antibiotic treatment has differing long- term impacts on the human throat and gut microbiome. PLoS One 5:e9836.

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761–2764.

Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale J V., DeBruyn JM, Lents NH, Oh J, Byrd AL, Park M, Kong HH, Segre JA, Bashan A, Gibson T, Friedman J, Carey V, Weiss S, Hohmann E, Jortha P, Turner K, Gumus P, Nizam N, Buduneli N, Whiteley M, Hyde E, Haarmann D, Lynne A, Bucheli S, Petrosino J, Metcalf JL, Parfrey LW, Gonzalez A, Lauber CL, Knights D, Ackermann G, Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S,

Tomberlin JK, Pechal JL, Crippen TL, Tarone AM, Lewis AJ, Tomberlin JK, Benbow ME, Carter DO, Metcalf JL, Bibat A, Knight R, Cobaugh KL, Schaeffer SM, DeBruyn JM, Finley SJ, Pechal JL, Benbow ME, Robertson BK, Javan GT, Hauther KA, Cobaugh KL, Jantz LM, Sparer TE, DeBruyn JM, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Guyer M, Michaud J-P, Gaétan M, Bishop C, Hill MO, Alpaydin E, Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Basak D, Srimanta P, Patranabis DC, Hoerl AE, Kennard RW, Liaw A, Matthew W, Saeys Y, Inza I, Larrañaga P, Guyon I, Elisseeff A, Metcalf JL, Xu ZZ, Weiss S, Lax S, Treuren W Van, Hyde ER. 2016. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS One 11:e0167370.

Joint Genome Institute. 2017. Integrated Microbial Genomes & Microbiome Samples. https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=ImgStatsOverview.

Kassinen A, Krogius-Kurikka L, Mäkivuokko H, Rinttilä T, Paulin L, Corander J, Malinen E, Apajalahti J, Palva A. 2007. The Fecal Microbiota of Irritable Bowel Syndrome Patients Differs Significantly From That of Healthy Subjects. Gastroenterology 133:24–33.

King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12C:128–135.

King JL, Wendt FR, Sun J, Budowle B. 2017. STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems. Forensic Sci Int Genet 29:21–28.

Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328–1333.

Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343–59.

Kodama Y, Mashima J, Kaminuma E, Gojobori T, Ogasawara O, Takagi T, Okubo K, Nakamura Y. 2012. The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. Nucleic Acids Res 40:D38–D42.

Kodama Y, Shumway M, Leinonen R. 2012. The sequence read archive: Explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56.

Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC, Komarow HD, NISC Comparative Sequence Program, Murray PR, Turner ML, Segre JA. 2012. Temporal shifts in the skin microbiome associated with disease flare and treatment in children with atopic dermatitis. Genome Res 22:850–859.

Krenke BE, Tereba A, Anderson SJ, Buel E, Culhane S, Finis CJ, Tomsey CS, Zachetti JM, Masibay A, Rabbach DR, Amiott E a, Sprecher CJ. 2002. Validation of a 16-locus fluorescent multiplex system. J Forensic Sci 47:773–85.

Kyrpides NC, Woyke T, Eisen JA, Garrity G, Lilburn TG, Beck BJ, Whitman WB, Hugenholtz P, Klenk H-P. 2014. Genomic Encyclopedia of Type Strains, Phase I: The one thousand microbial genomes (KMG-I) project. Stand Genomic Sci 9:1278–1296.

Kyrpides NC, Hugenholtz P, Eisen J a, Woyke T, Göker M, Parker CT, Amann R, Beck BJ, Chain PSG, Chun J, Colwell RR, Danchin A, Dawyndt P, Dedeurwaerdere T, DeLong EF, Detter JC, De Vos P, Donohue TJ, Dong X-Z, Ehrlich DS, Fraser C, Gibbs R, Gilbert J, Gilna P, Glöckner FO, Jansson JK, Keasling JD, Knight R, Labeda D, Lapidus A, Lee J-S, Li W-J, Ma J, Markowitz V, Moore ERB, Morrison M, Meyer F, Nelson KE, Ohkuma M, Ouzounis CA, Pace N, Parkhill J, Qin N, Rossello-Mora R, Sikorski J, Smith D, Sogin M, Stevens R, Stingl U, Suzuki K-I, Taylor D, Tiedje JM, Tindall B, Wagner M, Weinstock G, Weissenbach J, White O, Wang J, Zhang L, Zhou Y-G, Field D, Whitman WB, Garrity GM, Klenk H-P. 2014. Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. PLoS Biol 12:e1001920.

Lambert JA, John S, Sobe JD, Akins RA. 2013. Longitudinal analysis of vaginal microbiome dynamics in women with recurrent bacterial vaginosis: Recognition of the conversion process. PLoS One 8:e82599.

Langille MGI, Laird MR, Hsiao WWL, Chiu TA, Eisen JA, Brinkman FSL. 2012. MicrobeDB: a locally maintainable database of microbial genomic sequences. Bioinformatics 28:1947–1948. doi:10.1093/ bioinformatics/bts273.

Laurence M, Hatzis C, Brash DE. 2014. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One 9:e97876.

Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen J a, Gilbert J a. 2015. Forensic analysis of the microbiome of phones and shoes. Microbiome 3:21.

Leake SL, Pagni M, Falquet L, Taroni F, Greub G. 2016. The salivary microbiome for differentiating individuals: proof of principle. Microbes Infect 1–7.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997.

Li K, Bihan M, Methé BA. 2013. Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome. PLoS One 8:e63139.

Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 40:D115–D122. doi:10.1093/nar/gkr1044

Markowitz VM, Chen I-M a, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, Liolios K, Pagani I, Anderson I, Mavromatis K, Ivanova NN, Kyrpides

NC. 2012. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res 40:D123–D129.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10.

Mathieu A, Delmont TO, Vogel TM, Robe P, Nalin R, Simonet P. 2013. Life on Human Surfaces: Skin Metagenomics. PLoS One 8:e65288.

Meadow JF, Altrichter AE, Green JL. 2014. Mobile phones carry the personal microbiome of their owners. PeerJ 2:e447.

Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown G, Green JL, Bohannan BJ. 2015. Humans differ in their personal microbial cloud. PeerJ 3:e1258.

Metzker ML, Mindell DP, Liu X-M, Ptak RG, Gibbs RA, Hillis DM. 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proc Natl Acad Sci U S A 99:14292–14297.

Nakamura Y, Cochrane G, Karsch-Mizrachi I. 2013. The international nucleotide sequence database collaboration. Nucleic Acids Res. 41:D21–D24. doi:10.1093/nar/gks1084

Nayfach S, Pollard KS. 2015. Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria. bioRxiv DOI:10.1101/031757.

NCBI Genome. 2017. https://www.ncbi.nlm.nih.gov/genome/browse/. Date accessed: 2017-08-19.

Nierman W, Eisen JA, Fraser CM. 2000. Microbial genome sequencing 2000: new insights into physiology, evolution and expression analysis. Res. Microbiol. 151:79–84. doi:10.1016/S0923-2508(00)00125-X

Nishi E, Tashiro Y, Sakai K. 2014. Discrimination among individuals using terminal restriction fragment length polymorphism profiling of bacteria derived from forensic evidence. Int J Legal Med 129:425–433.

Nishi E, Watanabe K, Tashiro Y, Sakai K. 2017. Terminal restriction fragment length polymorphism profiling of bacterial flora derived from single human hair shafts can discriminate individuals. Leg Med 25:75–82.

Oh J, Conlan S, Polley EC, Segre JA, Kong HH. 2012. Shifts in human skin and nares microbiota of healthy children and adults. Genome Med 4:77.

Oh J, Byrd AL, Deming C, Conlan S, Kong HH, Segre JA. 2014. Biogeography and individuality shape function in the human skin metagenome. Nature 514:59–64.

Oh J, Byrd AL, Park M, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854–866.

Pechal JL, Crippen TL, Benbow ME, Tarone AM, Dowd S, Tomberlin JK. 2014. The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. Int J Legal Med 128:193–205.

QIAGEN. 2016. QIAGEN Multiplex PCR Plus Handbook. https://www.qiagen.com/.

Quagliarello B, Cespedes C, Miller M, Toro A, Vavagiakis P, Klein RS, Lowy FD. 2002. Strains of Staphylococcus aureus obtained from drug-use networks are closely linked. Clin Infect Dis 35:671–677.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

Reddy TBK, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, et al. 2015. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 43:D1099–D1106. doi:10.1093/nar/gku950.

Ross AA, Doxey AC, Neufeld JD. 2017. The Skin Microbiome of Cohabiting Couples. mSystems 2:e00043-17.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87.

Santiago-Rodriguez T, Cano R. 2016. Soil Microbial Forensics. Microbiol Spectr 1–15.

Savage DC. 1977. Microbial Ecology of the Gastrointestinal Tract. Annu Rev Microbiol 31:107–133.

Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. 2010. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proc Natl Acad Sci U S A 107:21242–21247.

Schmedes SE, King JL, Budowle B. 2015. Correcting Inconsistencies and Errors in Bacterial Genome Metadata Using an Automated Curation Tool in Excel (AutoCurE). Front Bioeng Biotechnol 3:138.

Schmedes SE, Sajantila A, Budowle B. 2016. Expansion of Microbial Forensics. J Clin Microbiol 54:1964–1974.

Schmedes SE, Woerner AE, Budowle B. 2017. Forensic human identification using skin microbiomes. Appl Environ Microbiol (In press).

Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods 13:435–438.

Schouls LM, Schot CS, Jacobs JA. 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 185:7241–7246.

Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. 2013. Computational meta'omics for microbial community studies. Mol Syst Biol 9:666.

Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. bioRxiv. doi: 10.1101/036103. http://biorxiv.org/content/early/2016/01/06/036103.abstract

Soergel DAW, Dey N, Knight R, Brenner SE. 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J 6:1440–1444.

Sommer MO, Dantas G, Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 325:1128–1131.

Song SJ, Lauber C, Costello EK, Lozupone C a, Humphrey G, Berg-Lyons D, Caporaso JG, Knights D, Clemente JC, Nakielny S, Gordon JI, Fierer N, Knight R. 2013. Cohabiting family members share microbiota with one another and with their dogs. Elife 2:e00458.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625–630.

Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res 43:D599–D605.

Tautz D. 1989. Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17:6463–6471.

Tibshirani R. 1996. Regression Shrinkage and Selection via the Lasso. J R Stat Soc 58:267–288.

Tilg H. 2010. Obesity, Metabolic Syndrome, and Microbiota Multiple Interactions. J Clin Gastroenterol 44:16–18.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12:902–903.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res 27:626–638.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1031.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. Nature 457:480–484.

Wang Y, Zhang Z, Ramanan N. 1997. The actinomycete Thermobispora bispora contains two distinct types of transcriptionally active 16S rRNA genes. J Bacteriol 179:3270–3276.

Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42:D581–D591. doi:10.1093/nar/gkt1099

Whitman WB, Woyke T, Klenk H-P, Zhou Y, Lilburn TG, Beck BJ, De Vos P, Vandamme P, Eisen JA, Garrity G, Hugenholtz P, Kyrpides NC. 2015. Genomic Encyclopedia of Bacterial and Archaeal Type Strains, Phase III: the genomes of soil and plant-associated and newly described type strains. Stand Genomic Sci 10:26.

Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wilke CO. 2016. cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2". R package version 0.7.0.

Williams DW, Gibson G. 2017. Individualization of pubic hair bacterial communities and the effects of storage time and temperature. Forensic Sci Int Genet 26:12–20.

Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 108:68–74.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462:1056–1060.

Yassour M, Vatanen T, Siljander H, Hämäläinen A, Härkönen T, Ryhänen SJ, Franzosa EA, Vlamakis H. 2016. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med 8:343ra81.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol Evol 8:28–36.