Sherier, Allison J. Improving Human Identification Using the Human Skin Microbiome. Doctor of Philosophy (Biomedical Sciences), January 2022, 133 pp., 16 illustrations, 3 tables, 48 bibliographies.

ABSTRACT

There are times when biological evidence has too low of quality or quantity of human DNA to provide enough information for human identification (HID). However, nucleic acids from the human skin microbiome are sources of genetic material that may be useful for HID. The studies in this dissertation test the hypothesis that specific single nucleotide polymorphisms (SNPs) of selected human skin microorganisms can be used to attribute an unknown microbiome sample to an individual.

The first study investigated how Wright's fixation index (F_{ST}) can be used to select potentially informative SNPs for HID. SNPs with high estimated F_{ST} were ascertained in three different ways to examine three distinct hypotheses. The hypotheses focused on testing whether a high F_{ST}, increased taxonomic abundance, and/or using a predetermined panel would be the most effective for HID. Classification accuracies ranged from 88 – 95%, and the method using the most taxa possible performed the best. Results from the study support that using genetic distance to select informative markers from the human skin microbiome for HID was viable. The predetermined panel only achieved an 88% accuracy, although it would be the most applicable of the tested method for forensic case work.

The second study focused on using F_{ST} estimations to select SNPs abundant in 51 individuals sampled at three body sites in triplicate for HID. The most common SNPs (present in $\geq$ 75% of the samples) which had F_{ST} estimates $\geq$ 0.1 were used with least absolute shrinkage and

selection operator (LASSO) to select a list of informative SNPs for HID. The final list (i.e., hidSkinPlex+) contains 365 SNPs and achieved a 95% classification accuracy on 459 samples. The hidSkinPlex+ lays the foundation for a targeted sequencing panel that can be used to further study the stability and specificity of human skin microorganism SNPs for HID applications.

KEYWORDS

Bacteria · Skin microbiome · Human identification · Microbial Forensics · Forensic profiling · Supervised learning · Machine learning · Targeted massively parallel sequencing · hidSkinPlex

IMPROVING HUMAN IDENTIFICATION USING

THE HUMAN SKIN MICROBIOME

Allison J. Sherier, B.S., M.S.

APPROVED:

_____

Dr. Bruce Budowle, Co-major Professor

_____

Dr. August Woerner, Co-major Professor

_____

Dr. Robert Luedtke, Committee Member

_____

Dr. Nicole Phillips, Committee Member

_____

Dr. Jerry Simecka, University Member

_____

Dr. Bruce Bunnell, Chair, Department of Microbiology, Immunology & Genetics

_____

Dr. Michael Mathis, Dean, Graduate School of Biomedical Science

IMPROVING HUMAN IDENTIFICATION USING THE HUMAN SKIN MICROBIOME

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

Allison J. Sherier, B.S., M.S.

Fort Worth, Texas

January 2022

ACKNOWLEDGEMENTS

Throughout my Ph.D. journey, I have received a tremendous amount of support and assistance. There are more people and more 'thank yous' than could ever be given in this written acknowledgment. The last six years were more challenging than I could have ever imagined but also filled with joy and accomplishments.

First, I would like to acknowledge my major professors, Dr. Bruce Budowle and Dr. August Woerner. Dr. Budowle has assisted me in growing as a researcher, scientist, writer, and presenter. Thank you for providing me with opportunities I had never imagined and constantly pushing me to be my best self. You always put your students first and adjusted your mentoring style to what I needed at the time. You have also taught me not to be so serious all the time and that jokes should have a place in almost all conversations. Dr. August Woerner, thank you for being one of the kindest and compassionate professors I have ever met. As a student and as a new mom, your constant encouragement to succeed means the world to me. Without your influence, I would not have learned how to code and found my true passion in research.

I also wish to express my deep appreciation for my advisory committee members, Drs. Nicole Phillips and Robert Luedtke, as well as my university member, Dr. Jerry Simecka. I have enjoyed working with each of you as you challenged me to grow and expand my knowledge.

To my mother, Mary Kay, who has always been my biggest supporter, you fought for my childhood education, then supported me in my fight to continue my education as an adult. You are a remarkable woman who has given me everything I needed to be a strong and independent woman. Mom, thank you for all the love and support you have given me. You have always been my number one fan, and I will always be yours.

To my husband, David, you have supported and loved me through all the crazy ups and downs of my schooling. You have been there day-in and day-out since my sophomore year of undergraduate studies and working through three degrees. Thank you for picking me up when I thought I was finished with academia and helping me find a way to combat burnout. Thank you for dealing with all my practice presentations (notably the oral qualifying exam when I put you to sleep multiple times with my practice answers). Thank you for taking on the workload of two people at home when I needed to give 120% of myself to my research.

To my son, Ambrose (A.J.), you came into this world with the spirit of curiosity that made me view the world differently. The last three years have been hard, but your joy for nature and learning science words has always put a smile on my face. Thank you for being patient with a mom working all hours of the day and reframing my thinking, so I focused on what is important in life.

To my uncle, Will, thank you for all the love and support of my scientific endeavors. I always love our conversations about science. Your breadth of knowledge astounds me, and one day, I hope to keep up with your sarcasm. Having you on my side has always made me proud and determined to make sure you are proud of me as well.

To my dad, Braxton, thank you for your unbridled enthusiasm for all of my passions. Thank you for always being willing to edit papers from undergraduate to Ph.D. I am so sorry we did not get to celebrate the end of my Ph.D. together; I know in my heart how proud you are of me and all that I have accomplished. I hope Memom, Papaw, Rocky, and you are having a party to celebrate for me and for what all three of you have helped me accomplish.

To my best friend, Amanda, we started this crazy journey of academia with each other nearly 12 years ago. I could have never imagined then that one of my biggest competitors would become my lifelong best friend. The path has not been smooth, and there were so many things

outside of our control, but we made it out -- alive -- together. I am looking forward to our next chapter outside of school.

To all the students, staff, and numerous visitors to the Center of Human Identification's Research and Development (R&D) lab that I have had the privilege to work with, I wouldn't be where I am today without each of you. Jonathan King, you have made being a part of the R&D lab feel like family. I will miss our walks to Ampersand to get coffee and the philosophical conversations that occurred during them. Thank you, Dr. Frank Wendt, for the many discussions about how to survive graduate school. Drs. Rachel Keiser and Nicole Novoroski, thank you for your endless support and teaching about wet benchwork and how to organize my research.

Finally, but not least, my first mentors were Kay Porter and Fara Williams, the generous and loving Louis Stokes Alliance for Minority Participation Program leaders. You both have spent countless hours encouraging, teaching, supporting, and loving me. Without you, I would have never known what graduate school was and how to accomplish my dreams.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

Supplemental Evidence for Human Identification

Skin Microbiome for Human Identification

Forensic genetics focuses on the identification of the source of unknown biological evidence from a crime scene using only nanograms or picograms of deoxyribonucleic acid (DNA). DNA profiling has been referred to as the 'gold standard in forensic science' (1), but DNA profiling does not always provide enough information to attribute evidence to a single individual. Identifying an unknown donor primarily involves comparing an evidence sample to a known reference sample with a defined set of short tandem repeats (STRs). STRs are microsatellites that consist of repeating nucleotide sequences or motifs and STRs are often targeted for human identification (HID; Figure 1) (2). The highly polymorphic nature of STRs allows for individualization of a person (or only a few persons) based on the difference in the number of copies of the repeating sequences. As the number of STR loci typed increases, so does the power of discrimination, where the power of discrimination is related to the probability of randomly selecting two people with the same STR profile. When a complete profile is obtained the discrimination power is extremely high.

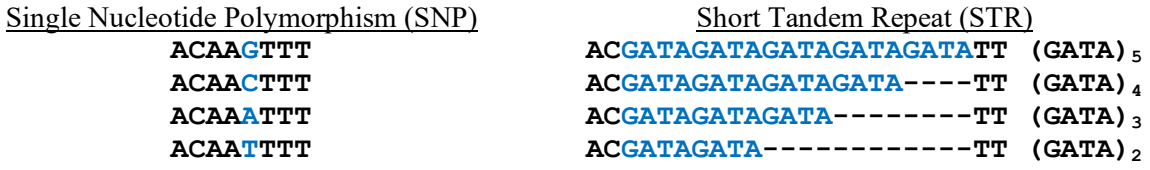
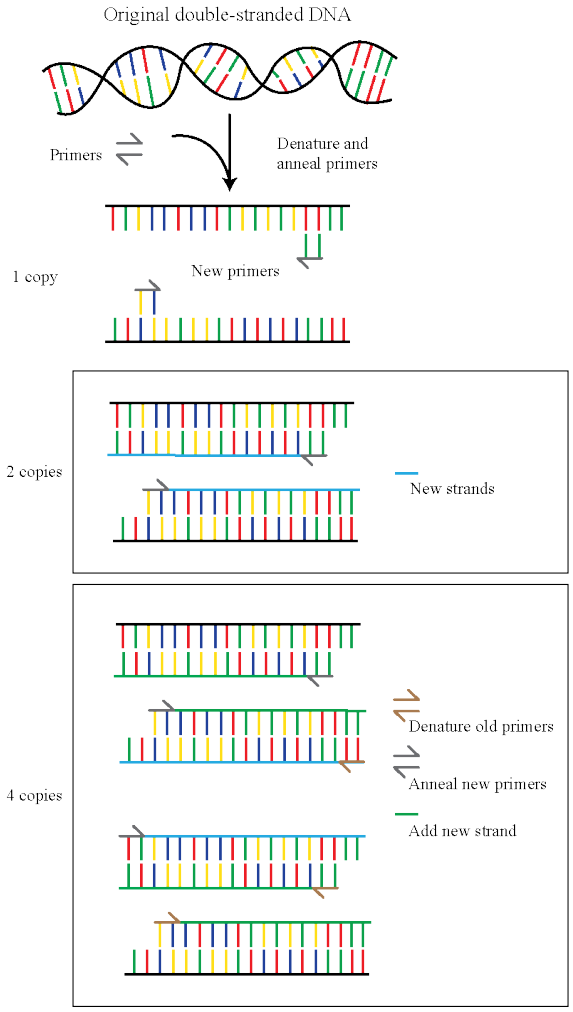


Figure 1. A comparison of a single nucleotide polymorphism (SNP) and a short tandem repeat (STR). On the left shows five different sequences with one nucleotide change at a SNP of interest (blue indicates differences). On the right demonstrates how a STR can have a range of allelic states (blue highlights repeating region).

Currently, 100 picograms (pg) (approximately the genomic equivalent of 16-17 cells) are considered the lower limit of input DNA needed to potentially obtain a complete profile for comparison of an unknown sample to a reference (2). However, in a forensic setting, biological

evidence can be highly degraded and/or have low amounts of DNA which can make it difficult to obtain a complete STR profile for HID. Even with 200 to 1,000 pg of input DNA, there are times when the DNA is fragmented or has lesions that prevent primers from binding and amplifying the STR of interest (2, 3). Primers are a short targeted single stranded DNA sequences that are used in the polymerase chain reaction (PCR; Figure 2). PCR is an enzymatic reaction that amplifies pieces of DNA. First, the original strand of DNA is denatured. Second, the complementary primers anneal with the targeted DNA. Third, amplification, i.e., copying, the strand of DNA occurs. A single cycle of PCR includes all three steps. The enzymatic reaction allows for a single copy or several copies of targeted DNA region to be replicated (i.e., amplified or copied) into millions of copies.



While many researchers are studying improved methods of detection and analysis, at times there may not be enough human DNA in biological evidence to obtain useful results. There are many laboratory-based approaches to attempt to improve the outcome of analysis of low quantity

and quality DNA, such as extraction and preparation methods including sample concentration, increased PCR cycle number, whole-genome amplification (prior to targeted amplification), post-PCR purification and increased injection times during capillary electrophoresis (CE) (for review see (4)). All of these methods, however, have shortcomings that include decreased reproducibility due to stochastic effects such as heterozygote imbalance and allele drop-in and drop-out, any of which can make interpretation difficult (Figure 3) (4).

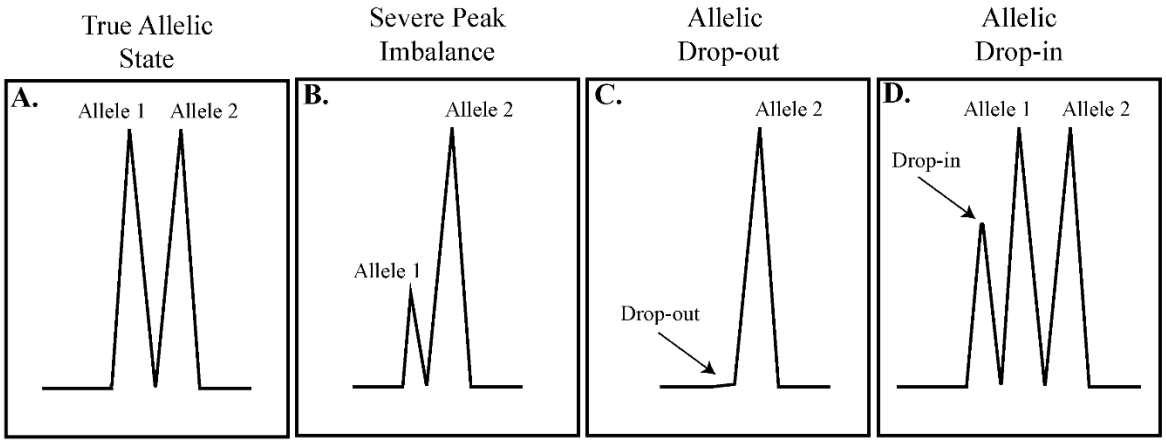


Figure 3. Examples of an ideal result and stochastic effects that can occur when low amounts of DNA are amplified with PCR. (A) Represents the ideal result for a heterozygote profile, i.e., two different alleles are apparent, where allele 1 and allele 2 are balanced peaks (i.e., similar heights). In the other 3 diagrams, stochastic variations are displayed which could result in an incorrect allele call if no additional information is available. Panel B is an example of an imbalance of two peaks resulting in uncertainty of whether it is heterozygote or a potential mixture. Panel C is an example of a heterozygote that appears as a homozygote due to allele drop-out (i.e., pseudo-homozygote). Panel D is an example of allele drop-in, with an allele present that is not from the donor.

There are other extra-nuclear sources of DNA to complement STRs for HID. One such source is mitochondrial DNA (mtDNA), which has a high copy number (several hundreds or thousands of copies) per cell (5). This increased copy number allows for easier detection of mtDNA than nDNA (nuclear DNA) in low quantity samples. While mtDNA provides a high copy

alternative to STR testing (i.e., nDNA markers), mtDNA has limited discrimination power because, unlike autosomal STRs, the entire mtDNA genome is a single haploid non-recombining marker. All individuals from the same maternal line have the same mtDNA profile, given that no *de novo* mutations have occurred in the timeframe. For example, in the absence of *de novo* mutations, a grandmother, mother, daughter, and their immediate male and female descendants all have the same mtDNA profile, making them indistinguishable based on this marker. There is a need for another source of DNA for HID that is high copy number and has high discriminatory power.

The human skin microbiome is a potential source for targeted DNA analysis that can enable the identification of a donor of biological evidence found at a crime scene using specific and sensitive markers of microorganisms that are shed from the skin (6-8). Being able to predict the contributor of a microbial profile shed from the skin allows for additional DNA evidence to support traditional DNA analysis methods for HID. It has been estimated that the human microbiome contains over 100 trillion microbes, which is 1:1 for all human cells and is perhaps ten times greater than the number of human nucleated cells (6, 7). An individual sheds approximately 30 microorganisms for every one squamous epithelial cell (9), therefore it is likely that people shed more microbial cells than human cells when they come in contact with items or other people (as suggested by Schmedes et al. (37)). The human microbiome is abundant, potentially allowing for targeted sequencing of key microorganisms of interest for HID in a forensic context.

Past Research in HID Using the Skin Microbiome

Initially, microbial genomics was not practical due to the challenge of sequencing microbial genomes. In 1995, Fleischmann et al. (11) sequenced the first whole microbial genome,

Haemophilus influenza. It took 13 months and cost almost one million dollars to sequence the 1.8 million base pairs (bp) of the *H. influenza* genome. By 2010 only 239 microbial genomes had been sequenced with 178 fully annotated microbial genomes being publicly available (12). However, by late 2021, over 19,000 complete microbial genomes had been sequenced and annotated with over 149,000 genomes either finished or in draft phase on Integrated Microbial Genomes and Microbiomes' website [1]. This substantial increase in the number of sequenced genomes is due in large part to the advent of massively parallel sequencing (MPS). MPS is a high throughput DNA (as well as ribonucleic acid (RNA)) sequencing technology that allows for many markers to be analyzed from multiple samples at one time. Currently, most of the microbial forensic community uses sequencing of the universal bacterial gene 16S ribosomal RNA (16S rRNA) gene for human and biological fluid identification from microbes present in the sample. Targeted 16S rRNA gene sequencing has continued to be used even with current day MPS advances because of relatively low cost and the substantial data on microbial taxonomic composition and phylogenetic diversity (13, 14).

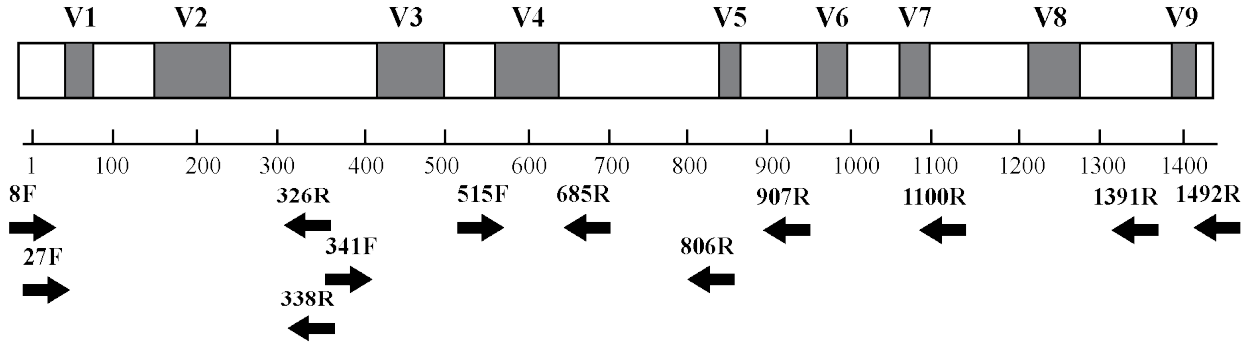


Figure 4. The 16S rRNA gene with commonly used forward (F) and reverse (R) primers. The greyed-out regions labeled V1 – V9 indicate the nine hypervariable regions in the 16S rRNA gene. Previous work focusing on the skin microbiomes for potential forensic applications have used a

[1] https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=ImgStatsOverview

variety of primers (represented by black arrows) to target specific areas of the 16S rRNA gene (8, 15-18).

Sequencing the Skin Microbiome

Many forensic studies have used 16S rRNA gene sequencing to investigate the taxonomic microbial differences of individuals for HID with mixed results. Although 16S rRNA gene sequencing is a standard marker for inferences on microbial taxonomy, there are several limitations to using 16S rRNA gene sequencing, or for that matter any single genetic marker, for certain high-resolution applications such as HID. These limitations include insufficient resolution at the species and strain level (19-21), copy number variation (22), inaccurate phylogenetic predictions (23), sample preparation bias (24-26) and PCR bias (27). An alternative to 16S rRNA gene sequencing is whole genome sequencing (WGS) which assays the entire genome of a microorganism and, in theory, allows for all the microorganisms in a sample to be sequenced (Figure 5) (11). While WGS may provide more comprehensive genome coverage and potential species/strain resolution, there are limitations with this approach as well, such as incomplete and stochastic coverage of the genome(s), differences in reliability and accuracy between sequencing platforms, preparation and analysis bias, lower sample throughput, and higher sequencing cost compared to 16S rRNA gene sequencing (28, 29). An alternative that may exploit the best features of 16S rRNA gene sequencing and WGS is targeted genome sequencing (TGS), which allows for coverage of more markers than 16S rRNA gene sequencing providing more information with less genome coverage than WGS (Figure 5). TGS (herein refers to a multiple, but limited number of specific targets) provides greater read depth and less stochastic effects than WGS for the targeted regions of interest. While TGS has limitations, as do all methods, its advantages are greater diversity than a

single marker system and more robust read depth for the targeted markers than WGS. Using TGS with a panel containing specific markers can increase prediction accuracy of unknown samples.

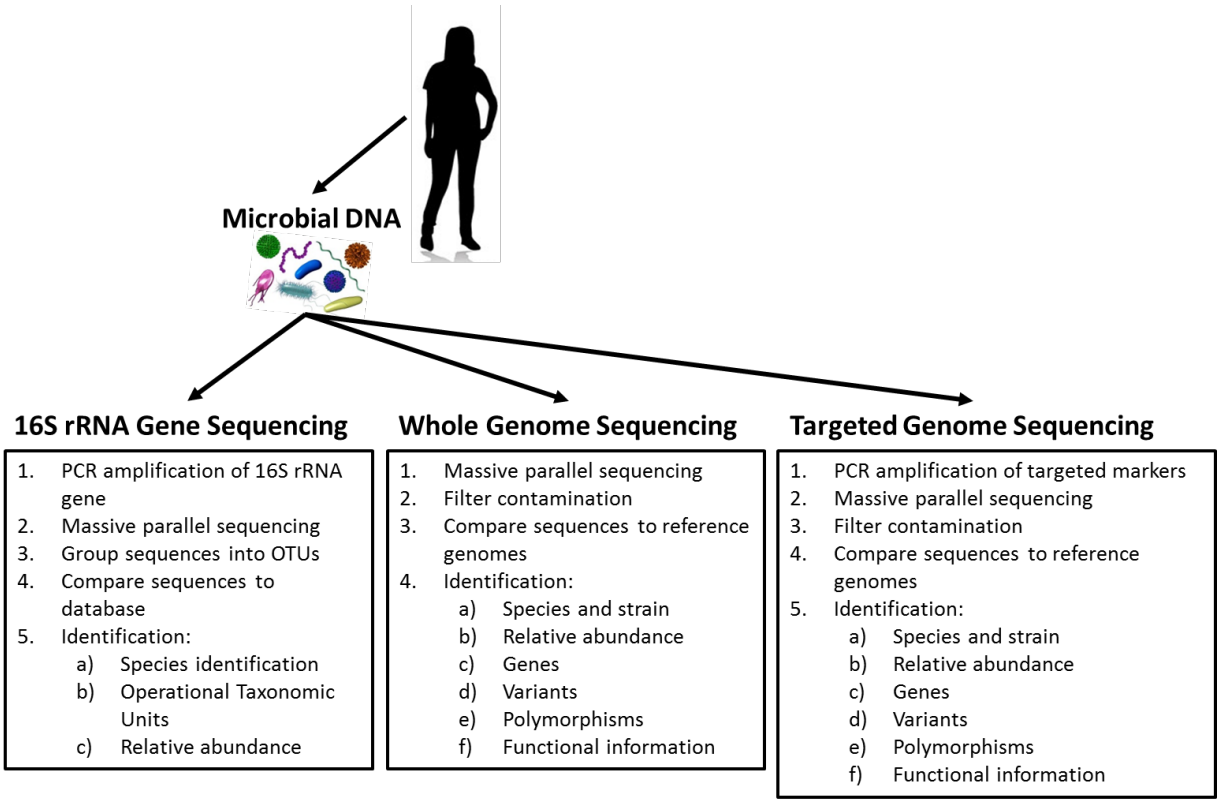


Figure 5. Microbial DNA can be characterized by 16S rRNA gene sequencing, whole genome sequencing (WGS), or targeted genome sequencing (TGS). The initial processing steps of a sample will be similar for all sequencing strategies. The choice of approach is dependent on the intended use of the microbial information and how it will be applied. Operational taxonomic units (OTU) for 16S rRNA gene sequencing are used to group closely related microorganisms in a sample, based on a similarity threshold (usually 97%). Microorganisms can also be classified by comparing the results of 16S rRNA gene sequencing to databases of known microorganisms.

Development of hidSkinPlex

The hypothesis that the human microbiome could be a useful target for forensics HID is supported to various degrees. Initial studies investigating the uniqueness of the human skin microbiome suggest that there are more unique species between individuals within a body site than within individuals between body sites (15, 17, 18, 30-33). Nonporous items/surfaces (phone,

keyboard, mouse, desk, etc.) handled by an individual can be linked back to the person that most often handles the item (8, 18, 31, 32, 34, 35). Additionally, species level differences between individuals have been observed by body location, suggesting that the skin microbiome may also be used to determine what part of the body encountered an item (33, 34, 36, 37). The resident taxa of the skin microbiome have been used to predict the human host for samples from nonporous surfaces and clothes an individual has worn (32, 34, 38-40).

When crimes occur, it can take days to weeks, if not longer, to obtain a reference sample from a person of interest for comparison to evidence collected from a crime scene. If there is an extended time between the crime where the unknown sample was deposited and the collection of a reference sample, the composition and genetic signatures may and likely will change to some degree. Thus, the stability of microbial markers over some lapsed time should be considered. Oh et al. (36) provided data on this parameter with the first shotgun whole genome metagenomics dataset with spatial and temporal sampling of the skin microbiome. The structure and stability of the skin microbiome were evaluated at 17 body sites at three different time points (up to three years), and the skin microbiome was found to be largely stable in that timeframe. Schmedes et al. (37) developed the hidSkinPlex based on (36), selecting markers that were stable and personally identifying. The hidSkinPlex was developed using two different taxonomic approaches: presence/absence or nucleotide diversity. Presence/absence was used to determine if a target region was present in a sample, and nucleotide diversity measured the strain-level heterogeneity of an individual's skin microbiome population. These two summary statistics were used to select specific regions of microbes that most contributed to the correct classification of an unknown sample to the individual from which it was collected (37). Using these taxonomic approaches Schmedes at al. (37) were able to design the hidSkinPlex which is a TGS sequencing panel that

contains 286 markers that are relatively stable and abundant on the human skin, providing a new avenue of genetic testing for HID.

Previous Methods for Human Identification with hidSkinPlex

The hidSkinPlex, a TGS panel, contains 286 selected markers (ranging from family, genus, and species level), which are contained in the microbial reference database of MetaPhlAn2 (37, 41). Schmedes et al. (37) designed the hidSkinPlex and used taxonomic approaches for HID. In (37), eight individuals were sampled at three body-sites in triplicate. Then the samples were sequenced with the hidSkinPlex. An average of 94% classification accuracy for host identification and 86% accuracy predicting the body site location of the sample were achieved (37). The classification accuracies for all samples collected (n=72; eight individuals, three body sites in triplicate), depending on body site and number of markers used for classification, ranged from 54.2 to 100%, indicating improvement is still needed to increase classification accuracy for the hidSkinPlex markers for HID (37).

Woerner et al. (10) used phylogenetic distance or genetic diversity to analyze samples collected from 51 individuals for three body sites in triplicate that had been sequenced with the hidSkinPlex panel. The classification accuracies varied from 53.6% on average for phylogenetic distance (patristic distance) and 71.7% on average for genetic diversity (Euclidean distance) (10). While the results of Woerner et al. (10) perhaps suggest that the patristic distance may be ill-suited for the application of HID, there are a variety of distance functions which may be better suited for attribution. Woerner et al. (10) demonstrated that a sample which was misclassified and attributed to a non-donor by both the phylogenetic and taxonomic methods could be better classified using Wright's fixation index (F_{ST}) (see below). Woerner et al. showed using F_{ST} in a misclassified

instance that there were many more markers with low F_{ST} values (≤ 0.2) between the sample and its misclassified source than between the same sample and its technical replicate. There were also less markers with high F_{ST} values (>0.2) between the same pairing, suggesting a different and perhaps more powerful way to predict the human host from microbial signatures.

Wright's Fixation Index

The fixation index, most commonly referred to as $F_{ST,}$ is one of the most common methods used to quantify genetic differentiation between populations (42). In the studies herein, F_{ST} was estimated from a sample of a single individual's skin microbiome (herein referred to as a population), as compared to another individual. F_{ST} reflects the probability that two alleles drawn randomly from a subpopulation are identical by descent (i.e., whether a segment of DNA shared by two or more microbes was inherited from a recent common ancestor) (Figure 6) (42). F_{ST} can be estimated by calculating the number of pairwise differences within populations compared to between populations (the population being the skin microbiome from an individual). Hudson et al. (43) proposed calculating F_{ST} as: $F_{ST} = 1-(H_w/H_b)$, where H_w is the mean number of pairwise differences within a population, and H_b is the mean number of pairwise differences between two populations (43).

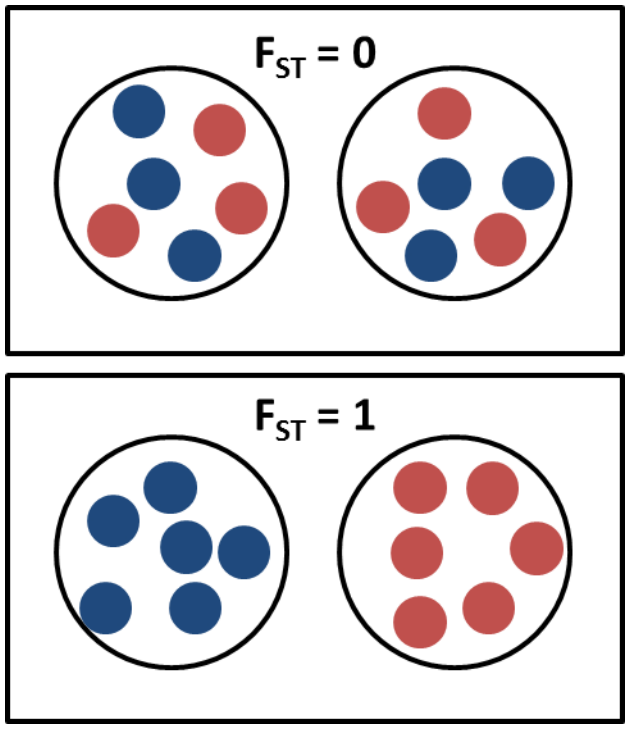


Figure 6. A cartoon depicting hypothetical F_{ST} values for two populations. The small blue and red circles represent alleles and the black circles around them represent distinct populations. Two populations showing extreme values of F_{ST} when alleles in the two populations have the same frequency (F_{ST}=0) or have no alleles in common (F_{ST}=1). F_{ST} values <0.05 may be considered little genetic differentiation while 0.05-0.15 may be considered moderate genetic differentiation (44). F_{ST} values > 0.15 may be referred to as a large differentiation between two populations (44, 45).

New Method to Select Markers from the Skin Microbiome for HID

F_{ST} has been used to select AIMs in human populations. AIMs have large differences in allele frequencies between human populations, and these differences have been used to estimate the ancestry of an individual (46). The same principle of AIMs may be applied to microbial HID. In this context, F_{ST} may provide insight into whether two microbial alleles are identical by descent and as such high F_{ST} microbial markers (e.g., SNPs) may provide information on whether or not two DNA samples show recent common ancestry (because they are derived from the same person) or if their similarities are likely due to chance. Additionally, SNPs allow for analysis of genetic differences between individuals' core stable microorganisms versus relying on the abundance of

specific taxa. With the development of a set of specific genetic markers, it may be possible to determine the frequency of certain alleles in the population allowing for a more precise identification of taxa present on an individual. If successful, selecting SNPs with high discriminatory power from the previously designed hidSkinPlex (37) panel should allow for the reduction of the number of markers needed, allowing for increased accuracy and decreased amplicon size. Additionally, the targeted panel would allow for more cost-effective investigation into the allele frequencies in different human populations (i.e., based on geographical location, lifestyle, and health) and determine the stability of the markers over time. With the addition of microbial profiling for HID there can be another source of DNA to accurately identify sources of biological evidence to support criminal investigations.

Machine Learning

Machine learning, also referred to as statistical learning, is a branch of computer science and mathematics wherein algorithms are trained to learn from data and identify patterns. There are two broad categories of machine learning, supervised and unsupervised learning. Supervised methods are used for prediction and are provided with input (independent variables) and output (dependent variables) data together, while unsupervised methods are provided with just input data (Figure 7). Supervised learning methods were investigated in this project, particularly support vector machine (SVM) and regularized logistic regression. SVM is a natural binary classifier where the algorithm creates a hyperplane (in two-dimensional space, a line) to separate the data into classes (Figure 8). SVM can be extended to a multiclass classification problem by using a one-versus-one approach wherein a binary SVM is learned for each pair of classes and predictions are combined to give the final predicted class. Another approach used in this study was regularized

logistic regression. Logistic regression is used to predict the probability and category of a dependent variable. Herein, the provided independent variables were the allele frequencies of SNPs as identified from select sites in the hidSkinPlex. Logistic regression can be regularized with the least absolute shrinkage and selection operator (LASSO, L_1 regularization) or the Ridge (L_2 regularization) to help reduce model complexity. Regularization in this context constrains the coefficients, reducing their sum of squares (in the case of Ridge) or reducing the sum of their absolute values (in the case of LASSO), and thus the model learned is simpler. When a model has less complexity it is also less likely to overfit, meaning that predictions on previously unseen data may tend to be more accurate. Additionally, LASSO tends to reduce the coefficients of uninformative independent variables to zero. As such, LASSO can be used simultaneously to perform prediction and to select a small number of markers to inform this prediction. In conjunction with LASSO, k-fold cross-validation (kCV) may be used while training algorithms. kCV leaves 1/k points out of the data set and then uses the held-out points to estimate how accurately the model predicts the expected outcome (Figure 9). All proposed machine learning methods have the potential to increase HID accuracy using the human skin microbiome. Specifically, regularized logistic regression has the ability to select a reduced number of SNPs needed for accurate attribution of a sample for further investigation of HID using the skin microbiome.

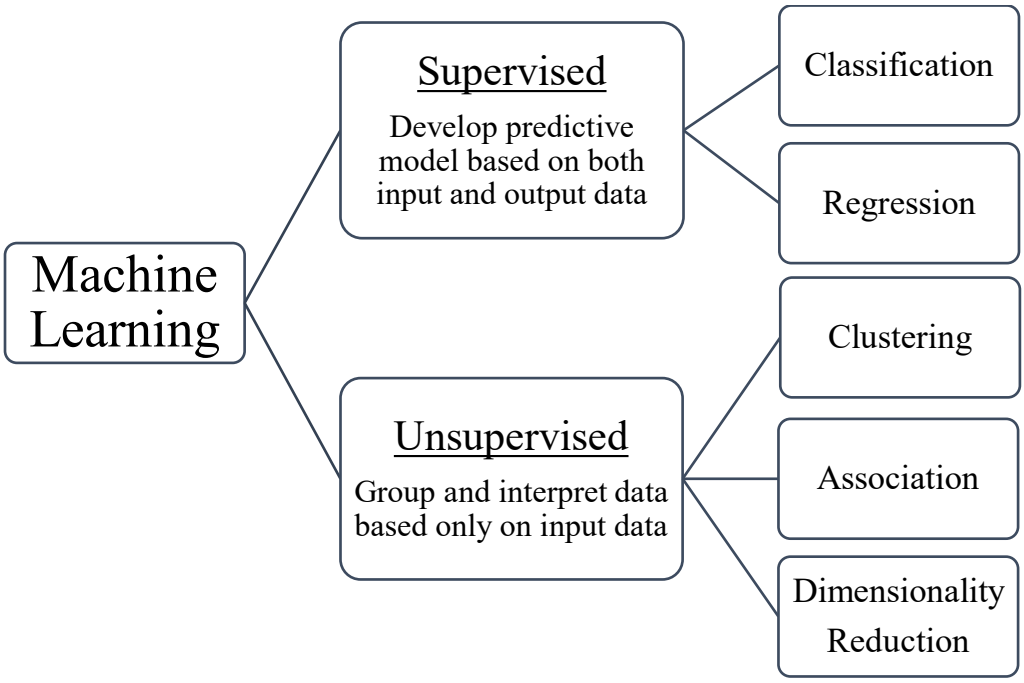


Figure 7. Flow chart showing the different categories of machine learning methods. Supervised learning uses training data with labels (herein, individuals) which can be used predict the outcome of future datasets. Unsupervised learning is a type of machine learning where the data are not labeled (i.e., the algorithm does not know the data associations) and the model tries to determine natural clusters or associations.

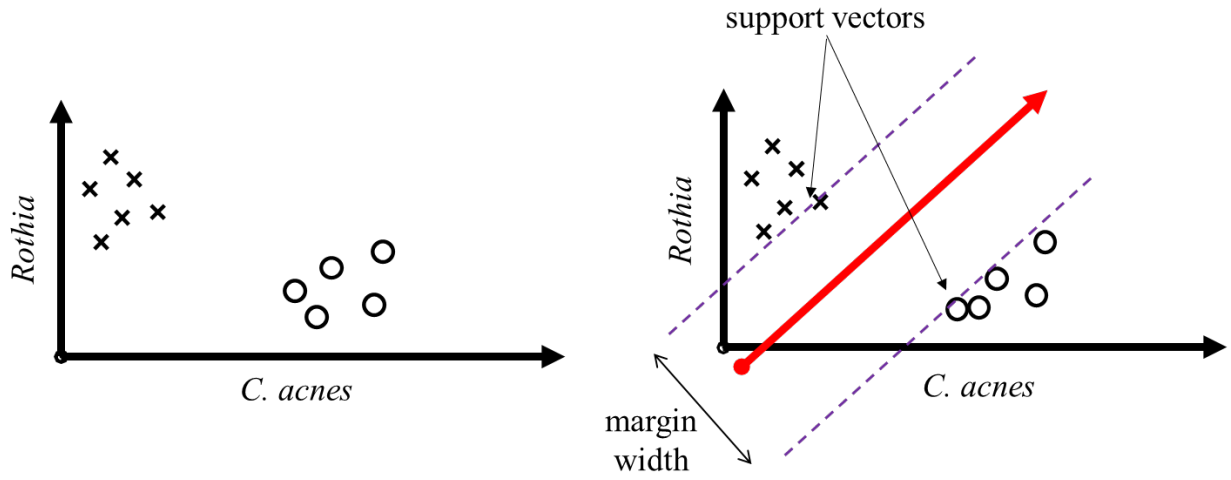


Figure 8. An example of a hard margin linear SVM. The algorithm finds the optimal hyperplane to separate data from two individual's independent variables (in the above, F_{ST} estimates for SNPs between the two individuals) (represented as X and O) in different classes. The SVM algorithm finds the points closest to the line for both classes, called support vectors. The distance between the line and the support vectors is calculated, called the margin. The objective of a hard margin linear SVM is to identify the hyperplane with the maximum margin width.

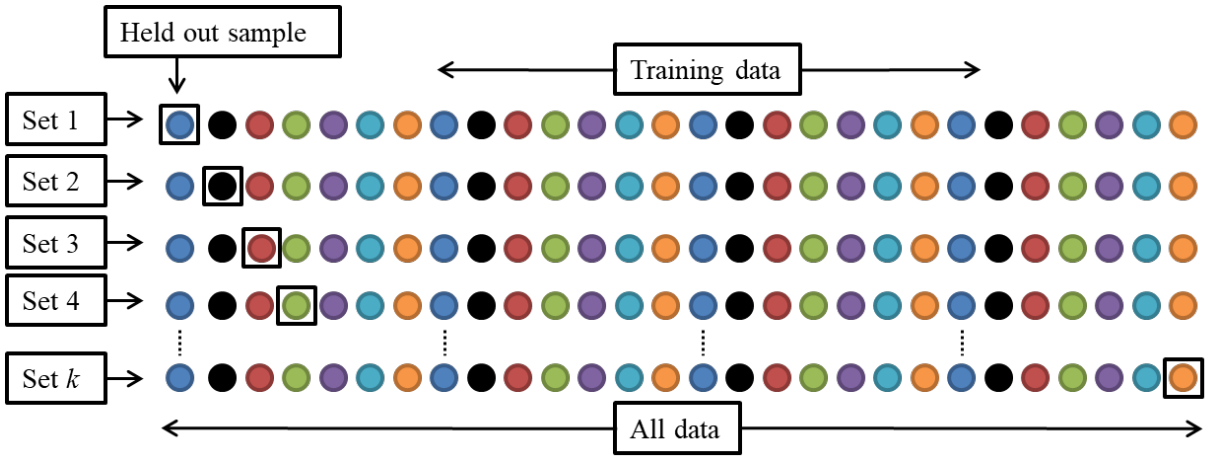


Figure 9. k-fold cross-validation (kCV) provides a statistical model with all the training data except one data point. Each round of kCV partitions the dataset into two subsets of data, one subset for training and one subset for validation (held out data). Held out samples are used for prediction and the remaining observations are used learn a predictive model. The procedure is then repeated by selecting a new sample and repeating the procedure until all observations have a prediction.

Research Question

While past research has provided valuable information for HID using the skin microbiome, using taxonomic differences between individuals has fallen short of consistently attributing an unknown sample to the person from which the sample was derived. Using genetic differences may provide more useful information to accurately identify the host of a skin microbiome sample. The goal for this proposed research project is to develop an improved targeted MPS panel which targets discriminatory SNPs from abundant microorganisms in the human skin microbiome for HID purposes.

This project will employ three technical replicates for three body sites per individual (n = 51 individuals). All possible comparisons of an individual's samples to every other sample in the data set was performed to determine which SNPs produce the highest classification accuracies when evaluated by supervised machine learning. The primary goals of this project are to use the hidSkinPlex data on 51 individuals to identify robust single nucleotide markers, reduce

misclassification rates, and lay the foundation for bioinformatic analyses of unidentified human skin microbiome samples.

In the dissertation herein, results and findings from two studies are described. **Chapter 2**, "Population informative markers selected using Wright's fixation index and machine learning improves human identification using the skin microbiome" (Sherier AJ, Woerner AE, Budowle B. 2021. Appl. Environ. Microbiol. 87:20), describes how leveraging genetic variants in stable microorganisms may provide a promising approach to microbial HID. This study used SVM to classify skin microbiome samples collected the non-dominant hand in triplicate from each of 51 individuals. Three methods for selecting SNPs with high F_{ST} estimates for classification of samples were performed. The first method, known as the *overall* method, selected a number of the highest-ranked SNPs (based on F_{ST} estimates) between two individuals using SVM for classification. The second method, known as the *per marker* method, focused on selecting several SNPs from each marker common between two individuals. The final method, known as the *selected* method, determined a single list of SNPs with high mean F_{ST} estimates that were common to most samples used in the training data set. Then SVM was used to determine the sample's human host. The resulting classifications from the SVM provided accuracies for each method. The *per marker* method had the highest accuracy at 95%, but the *overall* and *selected* methods still performed well at 92% and 88% accuracies, respectively. Determining a subset of SNPs contained within the hidSkinPlex that were successful lays the foundation for a redesigned targeted sequencing panel for HID.

Chapter 3, "Determining informative microbial single nucleotide polymorphisms for human identification" (Sherier AJ, Woerner AE, Budowle B. Submitted to Appl. Environ. Microbiol.), describes how LASSO was used to determine a reduced number of SNPs for

classifying skin microbiome samples to their hosts. Using the same 51 individuals sampled at three body sites (foot, manubrium, and hand) in triplicate, a reduced number of SNPs for potential HID was determined. A full list of nucleotide positions and their mean F_{ST} values were determined to select a reduced set of SNPs (F_{ST} mean estimate ≥ 0.1 and in greater than 75% of samples). LASSO was then used to determine a final list of 365 SNPs that could be used for human identification and provided 95% classification accuracy for 459 samples. Having a predetermined list of SNPs to use for HID provides an avenue for additional research into population allele frequencies and the overall stability of selected SNPs.

The studies comprising this body of work provide a new method for differentiating individuals based on informative SNPs from the skin microbiome. The resulting SNP panel can be used for studying the strengths and limitations of skin microbiome profiling for forensic HID. Future studies will focus on evaluating the new panel, referred to as hidSkinPlex+, on a larger sample size, assessing the stability of the SNPs over time, performance of the panel on mock case samples, and interpretation guidelines for using the reduced hidSkinPlex+. While the research contained in this dissertation focuses on using machine learning for HID, it may not be the desired approach for casework. After the markers in hidSkinPlex+ are more defined, traditional methods of interpretation for genetic data may be applied. More traditional methods may include using statistical analysis to determine the frequency of the genotype observed in the population allowing for probability (likelihood) calculations.

BIBLIOGRAPHY

1. Lynch M. 2003. God's signature: DNA profiling, the new gold standard in forensic science. Endeavour 27:93-7.
2. Butler JM. 2005. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, 2nd ed. Elsevier Academic Press, New York.
3. Alaeddini R, Walsh SJ, Abbas A. 2010. Forensic implications of genetic analyses from degraded DNA--a review. Forensic Sci Int Genet 4:148-57.
4. Budowle B, Eisenberg AJ, van Daal A. 2009. Validity of low copy number typing and applications to forensic science. Croat Med J 50:207-17.
5. Bogenhagen D, Clayton DA. 1974. The Number of Mitochondrial Deoxyribonucleic Acid Genomes in Mouse L and Human HeLa Cells: QUANTITATIVE ISOLATION OF MITOCHONDRIAL DEOXYRIBONUCLEIC ACID. Journal of Biological Chemistry 249:7991-7995.
6. Sender R, Fuchs S, Milo R. 2016. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell 164:337-40.
7. Sender R, Fuchs S, Milo R. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol 14:e1002533.
8. Neckovic A, van Oorschot RAH, Szkuta B, Durdle A. 2020. Investigation of direct and indirect transfer of microbiomes between individuals. Forensic Sci Int Genet 45:102212.
9. Percival SL, Emanuel C, Cutting KF, Williams DW. 2012. Microbiology of the skin and the role of biofilms in infection. International Wound Journal 9:14-32.

10. Woerner AE, Novroski NMM, Wendt FR, Ambers A, Wiley R, Schmedes SE, Budowle B. 2019. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci Int Genet 38:130-139.

11. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496-512.

12. Human Microbiome Jumpstart Reference Strains C, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, et al. 2010. A catalog of reference genomes from the human microbiome. Science 328:994-9.

13. Human Microbiome Project C. 2012. A framework for human microbiome research. Nature 486:215-21.

14. The Human Microbiome Project C, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC,

Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207.

15. 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 105:17994-9.

16. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Program NCS, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190-2.

17. Knight R, Metcalf JL, Gilbert JA, Carter DO. 2018. Evaluating the Skin Microbiome as Trace Evidence. National Criminal Justice Reference Service.

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

19. Bosshard PP, Zbinden R, Abels S, Boddinghaus B, Altwegg M, Bottger EC. 2006. 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. J Clin Microbiol 44:1359-66.

20. Mignard S, Flandrois JP. 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. J Microbiol Methods 67:574-81.

21. Fox GE, Wisotzkey JD, Jurtshuk P, Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166-70.

22. Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328-33.

23. Clayton RA, Sutton G, Hinkle PS, Jr., Bult C, Fields C. 1995. Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. Int J Syst Bacteriol 45:595-9.

24. Heikens E, Fleer A, Paauw A, Florijn A, Fluit AC. 2005. Comparison of genotypic and phenotypic methods for species-level identification of clinical isolates of coagulase-negative staphylococci. J Clin Microbiol 43:2286-90.

25. Tang YW, Ellis NM, Hopkins MK, Smith DH, Dodge DE, Persing DH. 1998. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. J Clin Microbiol 36:3674-9.

26. Woo PC, Ng KH, Lau SK, Yip KT, Fung AM, Leung KW, Tam DM, Que TL, Yuen KY. 2003. Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. J Clin Microbiol 41:1996-2001.

27. Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625-30.

28. Kwong JC, McCallum N, Sintchenko V, Howden BP. 2015. Whole genome sequencing in clinical and public health microbiology. Pathology 47:199-210.

29. Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL. 2017. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev 30:1015-1063.

30. 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 107:6477-81.

31. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, Brown GZ, Green JL, Bohannan BJ. 2014. Bacterial communities on classroom surfaces vary with human contact. Microbiome 2:7.

32. Kapono CA, Morton JT, Bouslimani A, Melnik AV, Orlinsky K, Knaan TL, Garg N, Vazquez-Baeza Y, Protsyuk I, Janssen S, Zhu Q, Alexandrov T, Smarr L, Knight R, Dorrestein PC. 2018. Creating a 3D microbial and chemical snapshot of a human habitat. Sci Rep 8:3669.

33. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930-8.

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

35. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S, Metcalf JL, Ursell LK, Vazquez-Baeza Y, Van Treuren W, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, Gilbert JA. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science 345:1048-52.

36. Oh J, Byrd AL, Park M, Program NCS, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854-66.

37. Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. 2018. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Sci Int Genet 32:50-61.

38. 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 126:815-23.

39. Lee S-Y, Woo S-K, Lee S-M, Eom Y-B. 2016. Forensic analysis using microbial community between skin bacteria and fabrics. Toxicology and Environmental Health Sciences 8:263-270.

40. Leung MHY, Tong X, Wilkins D, Cheung HHL, Lee PKH. 2018. Individual and household attributes influence the dynamics of the personal skin microbiota and its association network. Microbiome 6:26.

41. 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 12:902-3.

42. Wright S. 1951. The genetical structure of populations. Ann Eugen 15:323-54.

43. Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-9.

44. Hartl DL, Clark AG. 1997. Principles of Population Genetics. Sinauer Associates.

45. Banerjee AR. 2010. An Introduction to Conservation Genetics. The Yale Journal of Biology and Medicine 83:166-167.

46. Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B. 2016. Selection of highly informative SNP markers for population affiliation of major US populations. Int J Legal Med 130:341-52.

CHAPTER II

Population informative markers selected using Wright's fixation index and machine learning improves human identification using the skin microbiome

Appl Environ Microbiol 87:e0120821

Allison J. Sherier
August E. Woerner
Bruce Budowle

ABSTRACT Microbial DNA, shed from human skin, can be distinctive to its host and thus help individualize donors of forensic biological evidence. Previous studies have utilized single locus microbial DNA markers (e.g., 16S rRNA) to assess the presence/absence of personal microbiota in an effort to profile human hosts. However, since the taxonomic composition of the microbiome is in constant fluctuation, this approach may not be sufficiently robust for human identification (HID). Multi-marker approaches may be more robust. Additionally, genetic differentiation, rather than taxonomic distinction, may be more individualizing. To this end, the non-dominant hands of 51 individuals were sampled in triplicate (n = 153). They were analyzed for markers in the hidSkinPlex, a multiplex panel comprising candidate markers for skin microbiome profiling. Single nucleotide polymorphisms (SNPs) with the highest F_{ST} estimates were then selected for predicting donor identity using a support vector machine (SVM) learning model. Three different SNP selection criteria were employed: SNPs with the highest-ranking F_{ST} estimates 1) common between any two samples regardless of markers present (termed *overall*); 2) each marker common between samples (termed *per marker*); and 3) common to all samples used to train the SVM algorithm for HID (termed *selected*). The SNPs chosen based on criteria for *overall*, *per marker* and *selected* methods resulted in an identification accuracy of 92.00%, 94.77%, and 88.00%, respectively. The results support that estimates of F_{ST}, combined with SVM, can notably improve forensic HID via skin microbiome profiling.

IMPORTANCE There is a need for additional genetic information to help identify the source of biological evidence found at a crime scene. The human skin microbiome is a potentially abundant source of DNA that can enable the identification of a donor of biological evidence. With microbial profiling for human identification, there will be an additional source of DNA to identify individuals as well as to exclude individuals wrongly associated with biological evidence, thereby improving the utility of forensic DNA profiling to support criminal investigations.

Introduction

Determining the source of DNA evidence from a crime scene is the primary goal of forensic genetics. Identifying the molecular profile of a donor typically involves comparing short tandem repeat (STR) markers from an unknown sample(s) with a reference sample from a person(s) of interest. STRs are highly polymorphic, thus providing high powers of discrimination. However, forensic genetic evidence can often be degraded and/or contain low amounts of human DNA, making it difficult at times to obtain even a partial STR profile for human identification (HID). When an incomplete (or partial) STR profile is obtained, the discrimination power is reduced substantially. In such cases, there is a need for considering alternative approaches to assist in criminal investigations.

The human microbiome provides a promising alternative source of DNA that could supplement forensic human DNA analyses. Microbial cells outnumber their human counterparts by a ratio of 10:1 (though when considering all human cells, the ratio is estimated to be 1:1) (1, 2). Indeed, the skin microbiome is an abundant source of microbes, with an estimated ~10,000 bacteria/cm^2 (3). In contrast, human nuclear DNA (nDNA) is far less abundant on a per copy basis. For example, Schmedes et al. (4) swabbing a similar area of the skin obtained a quantity of human DNA that was equivalent to four diploid cells. In contrast, the DNA of the human skin microbiome from the same extract provided sufficient information for identification of the donor of the sample (4).

The 16S ribosomal RNA (rRNA) marker has traditionally been used in the context of human microbiome profiling. The human skin microbiome has been characterized for multiple individuals and multiple body sites using 16S rRNA sequencing demonstrating that the human skin microbiome is a potential source of trace evidence (5-14), but there still is need for

improvement. These studies have focused on the taxonomic diversity of specific microbial species to determine the relationship between an unknown sample and its potential donor. However, previous studies have had varying success rates for HID and were typically based on a small number of samples (i.e., < 15 individuals) (15-18). The limited success of these investigations could be attributed to their reliance on the presence/absence (or quantitation) of specific microbes as evidence for a "match" between an unknown sample and a reference sample. Environmental interactions and temporal shifts are common phenomena in microbiomes (19). Specifically, microbes from the skin can also be shared and exchanged between cohabiting and non-cohabitating individuals when they come in contact with each other or items (9, 20-22). Moreover, several studies have also claimed that 16S rRNA lacks the necessary phylogenetic resolution for HID (6, 9, 16, 23-27). All the above suggest that the taxonomic and phylogenetic constitution of microbiome is in constant fluctuation, and that using presence/absence of specific microbial taxa as evidence of a match could be limiting or possibly misleading.

However, a better system possibly consists of identifying discriminatory skin microbial features in which stability decays minimally over time. Consequently, recent work has focused on targeting a number of stable taxon-specific markers to improve accuracy of HID (4, 28, 29). Oh et al. (28) completed one of the first whole genome sequence studies of the human skin microbiome for multiple body sites, providing detailed information about abundant and stable microorganisms. The hidSkinPlex (4), for example, is a multiplex panel based on the data of Oh et al. (28) and includes 286 markers, ranging from the level of the genus to subspecies of 22 different microbial clades. The markers were selected based on their abundance and temporal stability (up to three years) as well as their prevalence across body sites (4, 28). Using specific stable markers with a wide phylogenetic range allows for the selection of specific features from the skin microbiome

that may improve HID. For example, the markers chosen by Schmedes et al. (4) were able to achieve accuracies with a range of 54.00% - 100.00% using presence/absence and nucleotide diversity with two machine learning methods, albeit with a limited sample size.

Ancestry informative markers (AIM) regularly used in human bio-ancestry studies commonly have high F_{ST} estimates (30, 31), wherein a few high F_{ST} markers are first mined from genomes and then used to predict population groups. A promising approach to identifying human hosts could use measures of genetic differentiation, specifically the F-statistics (for example the Fixation Index also known as F_{ST}) (32) for assessing microbial populations. F_{ST} can be estimated by evaluating orthologous SNPs in two different skin microbiome populations (i.e., skin microbiome samples from different individuals). F_{ST} estimates could provide insight into whether the alleles of a marker observed between microbial populations are identical by descent, allowing for better discrimination between microbial populations, which in turn may improve the accuracy of associating a skin microbiome sample with its respective human host.

Previously Woerner et al. (29) estimated F_{ST} values between two sample populations: a sample that was incorrectly associated with another host. Their work showed that even though the central value (i.e., mean) F_{ST} would also lead to an incorrect classification, the use of high F_{ST} SNPs would lead to the correct classification. However, the Woerner et al. study was only a proof-of-concept because only two samples were analyzed, and classification of the hosts based on the F_{ST} estimations was not performed. In this current study, a novel approach to accurately associate skin microbiota to their respective hosts is described. The non-dominant hands of 51 individuals were sampled in triplicate, and the DNA was analyzed using the hidSkinPlex panel. F_{ST} estimates were then computed using SNPs found across the sequenced markers to assess genetic differentiation between inter-and intra-individual microbiome populations. A select number of

SNPs displaying the highest F_{ST} estimates were chosen applying three different approaches: those with the highest-ranking F_{ST} estimates 1) common between any two samples regardless of taxonomy (termed *overall*); 2) per common marker between samples (forcing a more uniform distribution on taxonomy, termed *per marker*); and 3) markers common to all samples that are used to train the subsequent machine learning algorithm (termed *selected*). Each approach focused on a specific hypothesis to determine if using the overall highest-ranking SNPs, maximizing taxa, or a common selected panel could increase classification accuracy of unknown skin microbiome samples. These SNPs were used as data points for classification by a support vector machine (SVM) learning approach. The predictive capabilities of the SVM to match samples to their human hosts were compared across all three methods of SNP selection.

Results

F_{ST} estimations for skin microbiome samples. As previously, described in Woerner et al. (29), 51 individuals' non-dominant hands were sampled in triplicate and analyzed for the markers in the hidSkinPlex panel. The samples were split into training (n = 26 individuals in triplicate) and test data (n = 25 individuals in triplicate) sets. A total of ~69 million quality-controlled reads with a mean of 893,355 (SD = 362,436) per sample remained after read preprocessing for the training set. The test data set had a total of ~72 million mapped reads with an average of 964,161 (SD = 418,058) mapped reads per sample. F_{ST} was estimated over all pairs of individuals for every orthologous nucleotide in the hidSkinPlex within the training and test data sets.

After estimating F_{ST} for all pairwise comparisons that had at least 1x read coverage, the average number of nucleotides with an F_{ST} estimate greater than zero for each pairwise comparison

in the training data set was 24,809 (SD = 8,502; 2,590 min to 52,459 max) (Table S1). The test data set had a mean of 22,789 (SD = 9,657) for single nucleotide positions with a F_{ST} estimate greater than zero. When analyzing F_{ST} estimates for all pairs, only 236 markers of the 286 markers in the hidSkinPlex were seen in at least two samples being compared from the training data set. As a reminder, each marker in the hidSkinPlex is associated with some level of microbial taxonomy (e.g., stably present in *Cutibacterium acnes* at the species level). The reduced number of markers were only from eight species and one family. With one species, *Corynebacterium pseudogenitalium*, only seen in one comparison of two samples with both samples collected from the same individual (Table S2).

SVM analysis of training data set. SVMs are natural binary classifiers, and for the purposes of this study, each person is considered as a separate class. SVMs can be extended to multiclass classification by using one-versus-one (OvO) decomposition, wherein a classifier is built for each pair of classes (individuals). OvO classifiers were created using SNPs, selected based on high-ranking F_{ST} estimates, specific to the pair of individuals. The multi-class classification was estimated by using a simple tally of votes (see Methods). Parameter optimization included varying the number of SNPs, the minimum number of reads and the SVM cost (C), and the best combination of parameters were identified for each SNP selection method. The best combination was selected from the training data based on classification accuracy with a tie-breaking rule using the mean prediction accuracy.

The three methods of selecting SNPs with the highest-ranking F_{ST} estimates were termed *overall*, *per marker*, and *selected*. While all three methods focused on the SNPs with the highest-ranking F_{ST} estimations, each method varied on the number of markers and SNPs used to classify an unknown sample. The variation in the three methods were developed to answer distinct

hypotheses about how SNP selection methods affect HID and to determine which method had the highest accuracy, as assessed in the test data set. The *overall* method tested whether accuracies can be increased by selecting the highest-ranking SNPs, regardless of the markers present. The *overall* method selected SNPs with the highest-ranking F_{ST} estimates in each pair of samples (although it could lead to less diverse distribution of taxa). The *per marker* method tested whether maximizing the number of taxa used for classification could increase classification accuracy, even if doing so relied on SNPs with lower F_{ST} estimates. The *per marker* method selected the SNPs with the highest-ranking F_{ST} in each orthologous marker in a pair of samples. The *selected* method tested whether using SNPs that were common to all samples in the training data, used to train the SVM, could be used to increase accuracy of identification. The *selected* method relied on a predetermined number of common SNPs, which had high-ranking F_{ST} estimations for all comparisons in the training data set. Each selection method was then compared under different parameter values (i.e., the number of SNPs, minimum sequence reads, and SVM cost) using a customized SVM approach designed specifically for HID (see Methods).

Overall Method. The *overall* F_{ST} selection focused on choosing the highest F_{ST} SNPs for each pairwise comparison (i.e., 500, 1,000, or 2,000; note with this method that some of the highest-ranking SNPs had F_{ST} estimates close to zero). The number of selected high-ranking SNPs was tested with all possible combinations of minimum reads and SVM cost (i.e., the C hyperparameter). The data training set compared the accuracy of 75 parameter combinations, and 12 combinations performed best, classifying 76 out of 78 samples correctly, yielding a 97.44% accuracy (Table S3A). Using the highest prediction probability to break the tie of the 12 options, the 500 SNPs with the highest F_{ST} estimations, minimum read depth of 250, and SVM cost of 1 were the optimal parameters. The two incorrectly classified samples were S028_R3 and S029_R2

(Figure 1), which had some of the lowest number of markers (mean ± SD) (S028_R3 120.80 ± 0.47, S029_R2 152.00 ± 11.91) and SNPs (S028_R3 823.50 ± 131.28; S029_R2 1059.00 ± 144.45) for analysis among the training set samples. The mean number of markers for the *overall* method with the training data set was 146.20 (SD = 15.72), and the mean number of SNPs was 1,036.00 (SD = 160.25). The mean number of taxa seen was only 3.89 (SD = 0.86).

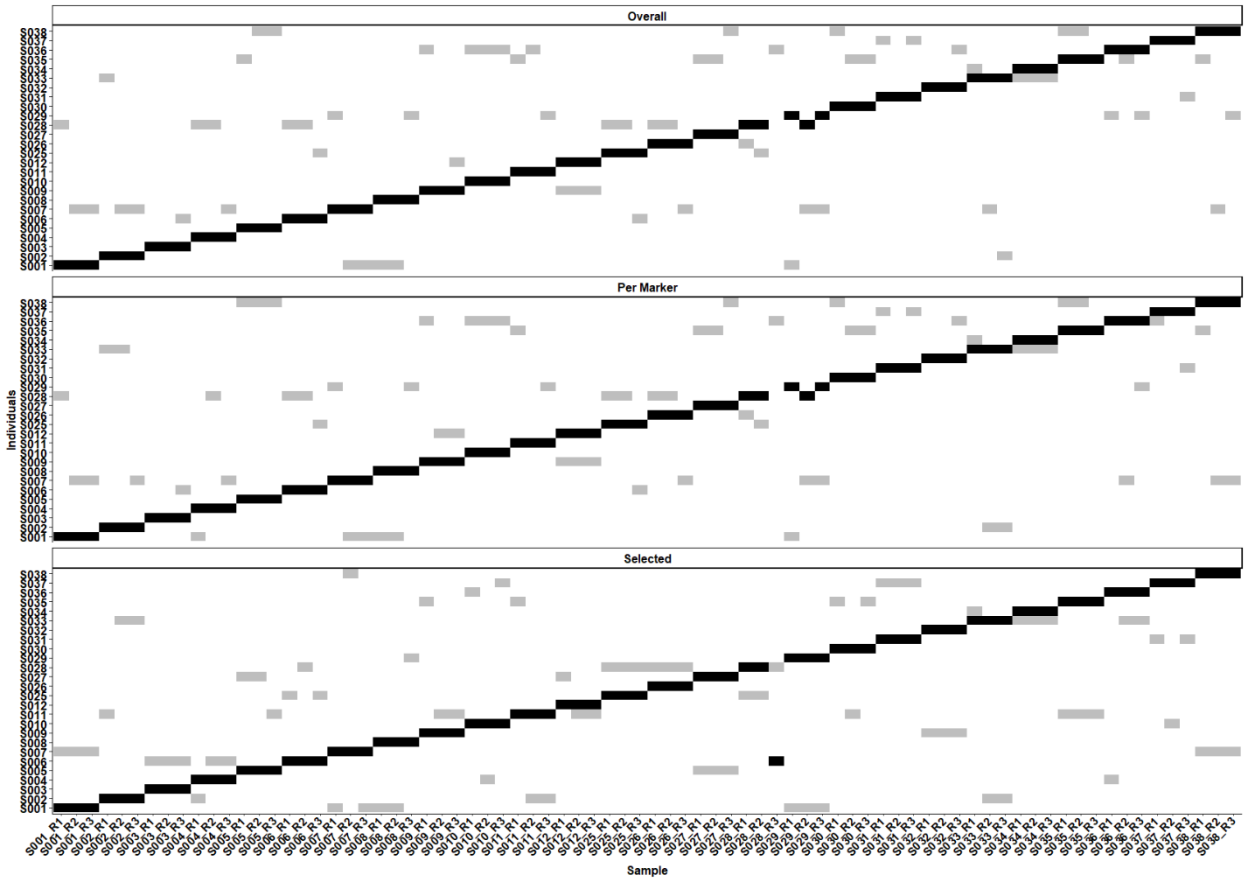


Figure 1. Training data set matrices showing rank #1 (black) and #2 (gray) for classification. The three matrices are labeled with the nucleotide selection method (i.e., *per marker, overall,* or *selected)* used at the top of the individual graphs. The three selection methods chose SNPs with the highest-ranking F_{ST} estimates. The *overall* method optimized 250 SNPs for the pairwise comparison, *per marker* method optimized 5 SNPs per marker, and *selected* had a set of 150 SNPs that were common in the training data set. The x-axis lists all samples with the individual number and replicates (S0## = individual number, R# = replicate number). The y-axis lists the possible groups, i.e., individuals, a sample could be classified.

Applying the optimized parameters to the test data set (n = 25 samples in triplicate) yielded a classification accuracy of 92.00% (69/75) with classification error of the model likely between 2.99% and 16.60% with 95% confidence (R package exactci (33)) (Figure 2). The test data set for the *overall* method assayed a larger number of markers (152.10 ± 14.16)) but had fewer SNPs (1,026.00 ± 166.89)) on average when compared to the training set. While six samples were incorrectly classified, four of the incorrect classifications involved S014 and S042. The other two incorrectly classified samples S044 and S046 ranked as #1 (Figure 2).

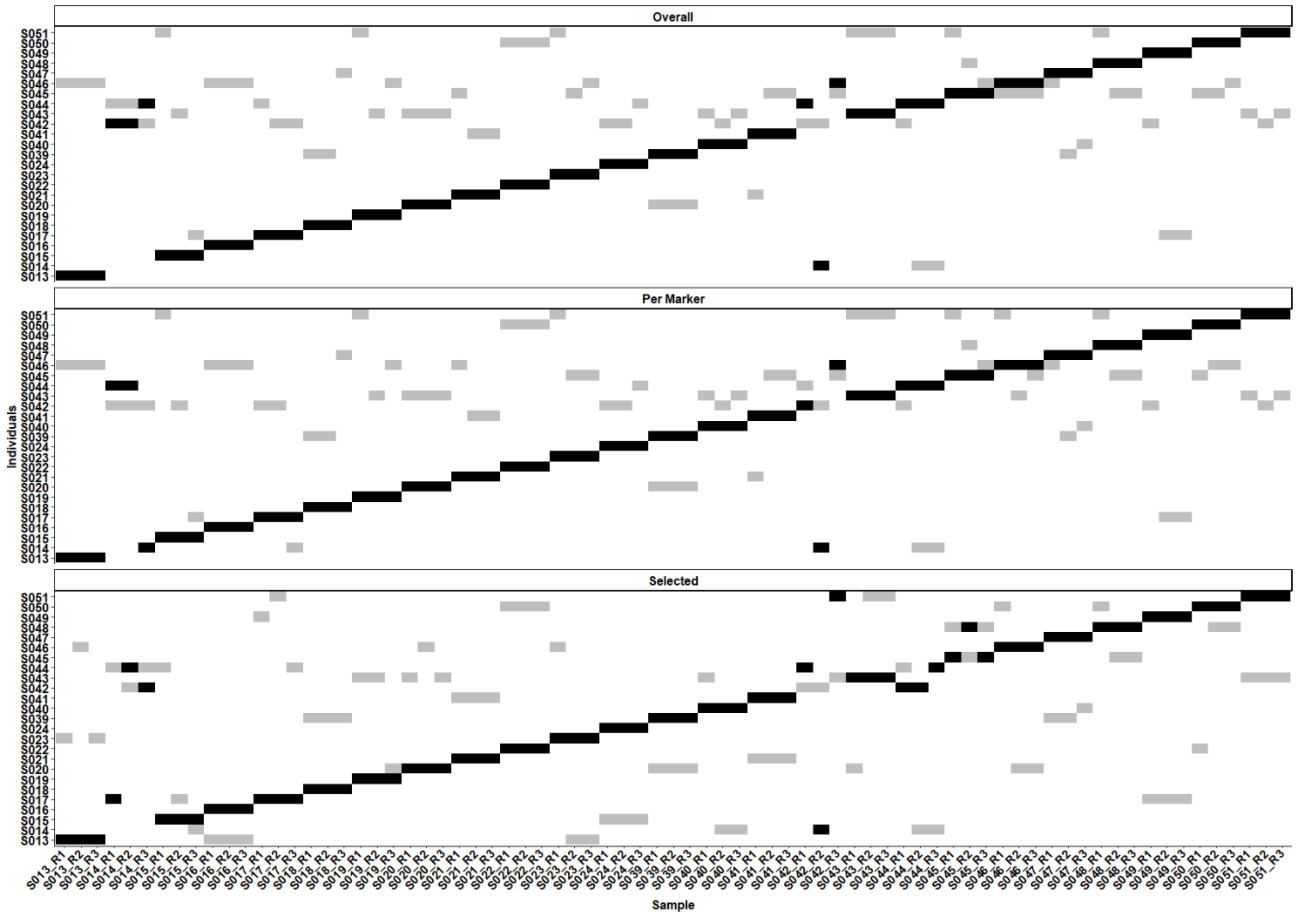


Figure 2. Test data set matrices showing rank #1(black) and #2 (gray) for classification of samples for the three methods of selecting the highest-ranking SNPs based on their F_{ST} estimation. The top matrix is the overall method, which chose the 250 highest SNPs in any given pairwise comparison. The second matrix shows the per marker method using training set optimized parameters of 5 SNPs per marker in a pairwise comparison. The bottom matrix shows the selected method that had 150 prechosen SNPs that were common and had the highest-ranking F_{ST} estimates in the training data set.

Per Marker Method. The *per marker* method focused on the highest-ranking F_{ST} estimates within each marker to achieve the largest taxonomic diversity possible. With the *per marker* approach up to a specified number of SNPs with the highest-ranking F_{ST} estimates (i.e., 5, 10, or 25) were selected per orthologous marker in a pair of samples. The *per marker* approach allowed for the widest variety of taxa (5.11 ± 1.15) and the largest number of markers (151.40 ± 27.98) to be used for classification with the training data. There was a total of 75 parameter combinations, and 11 parameter combinations provided the same highest prediction accuracy (Table S3B). Using 5 SNPs per marker with a minimum read depth threshold of 250 and a SVM cost of 10 yielded the highest accuracy and the highest mean confidence prediction. Each SVM analysis for the optimized parameters for the *per marker* method had a mean of 1,650.00 SNPs per SVM classification (SD = 309.15). The *per marker* training set generated a 97.44% accuracy with only two misclassifications out of 78 samples. S028_R3 and S029_R2 were also incorrectly classified samples with the *overall* method (Figure 1).

Using the optimized parameters, five SNPs with the highest-ranking F_{ST} estimates per marker, 250 read minimum, and a SVM cost of 10, the test data set produced an accuracy of 94.70%, with classification error between 1.47% to 13.11% (binomial 95% confidence interval). Four samples were classified incorrectly out of 75 (Figure 2). All four comparisons were from two samples, S014_R1/R2 and S042_R2/R3. One sample in the test data set, S014_R3, had a three-way tie, based on votes, with three potential candidates, S014, S042, and S044. S014 was ranked #1 of potential candidates because it had the highest mean prediction accuracy out of the three possible choices. All three replicates for S014 had S044 ranked #1 or #2. Additionally, S044 was classified correctly, but it had a close association with S014 and S042 with those two classes

ranked #2 and #3 for S044. Having the same classes ranked highly for S014, S042, and S044 is of particular interest because S042_R2 was classified as S014, indicating that they could have potentially arisen from the same host, a potential sample mix-up, or close relationship of the hosts.

Selected Method. The *selected* method used predetermined SNPs for analysis (i.e., 50 to 2,000). The number of SNPs were selected based on the number of the markers in the hidSkinPlex and their base pair length. The different number of SNPs chosen for the *selected* method was determined based on the maximum number of SNPs (~2,000) used in the previous two methods that had the highest classification accuracy. Out of 175 parameter combinations, ten combinations yielded the highest accuracy of 98.70%. The optimized parameters for *selected* F_{ST} were 150 high F_{ST} SNPs, a minimum of 500 reads, and a SVM cost of 1,000 (Table S3C). The 150-common SNPs represented 22 markers, one family and two species (*Propionibacteriaceae, Cutibacterium acnes*, and *Cutibacterium humerusii*) from the hidSkinPlex. The *selected* method had a training accuracy of 98.70%, with only one sample incorrectly classified. The incorrectly classified sample, S028_R3, was also incorrectly identified with the other two selection methods. The difference in the *selected* method was that S028_R3 ranked #2 based on its votes, and S006 was ranked as #1 by votes (Figure 1) and was a notable change in the rank of the correct group classification for S028_R3 which changed from rank ten (in *per marker* and *overall* methods) to rank two. In the training data set, S028 R3 had a mean of 43.94 markers (SD = 0.42) for all possible pair of comparisons.

When the test data set was evaluated with the parameters of 150 SNPs with the highest-ranking F_{ST} estimates from the training data set, 500 minimum reads, and a cost of 1,000 for the selected F_{ST} method, the accuracy decreased to 88.00% with a classification error of 5.63% to 21.56%. Only 66 out of 75 samples were correctly classified. Of the 11 incorrectly classified

samples: three belonged to S014, three to S042, two to S017, two to S044, and one to S045 (Figure 2). Individuals S042 and S045 did not have as many SNPs in their replicates, 146.90 ± 4.33 and 140.10 ± 1.85 across all replicates, respectively, when compared to other individuals, which may have impacted classification. However, missing data alone cannot explain the decreased accuracy with the *selected* method as other replicate sample pairs did not contain all 150 specified SNPs and were classified correctly.

For the training data set the difference in methods and parameters only resulted in 1.26% (or one sample) difference in classification accuracy, but the goal of the training data set is to find the parameter combinations that result in the highest accuracy. Testing the optimized parameters on the test data set provides a better indication of how the method and optimized parameters perform on unknown data. The classification accuracy results of the three methods ranged from 88.00% - 94.00% but accuracy rates are not significantly different (McNemar's chi-square test, Table S4). All three methods had issues determining the correct classification for samples from S014 and S042, but the *selected* method also had difficulty correctly associating S017, S044, and S045. While the *selected* method performed better than the other methods on the training data, it was the method predicted to most likely be overfit due to SNPs being chosen based on their presence in the training data set.

Study of Misclassified Sample. In this study, a misclassification was considered any unknown sample that was assigned to an incorrect individual. In essence, this error assumes the analysis achieves uniqueness which may not be realistic with these data. Thus, a misclassification may not be a true error. More studies with refined markers/SNPs and larger sample sizes are needed to determine the host resolution of the system. Plausible explanations were sought as to why one sample was consistently misclassified in the training data set before the optimized

parameters were used with the test data set. In the *overall* and *selected* methods, S028_R3 was classified as S036 (full rankings Table S5). S028_R3, which only had a total of 16 votes (*per marker* and *overall* methods), out of the potential 25 votes, ranking it number ten on the list of potential donors, was the only sample in the training data set that did not have the actual contributor ranked in the top three of potential candidates. For the *selected* method, S028_R3 was ranked #2 and had 24 votes, while S006 was ranked #1. When compared to S036, S028_R3 had much lower read coverage S036 for the markers that are orthologous between samples. S028_R3 had the fewest markers in common when estimating F_{ST} compared to any other sample that was analyzed. The reduced amount of data available for classification may be associated with individual S028_R3 having low read depth coverage or no reads for the SNPs of interest (Figure 3B).

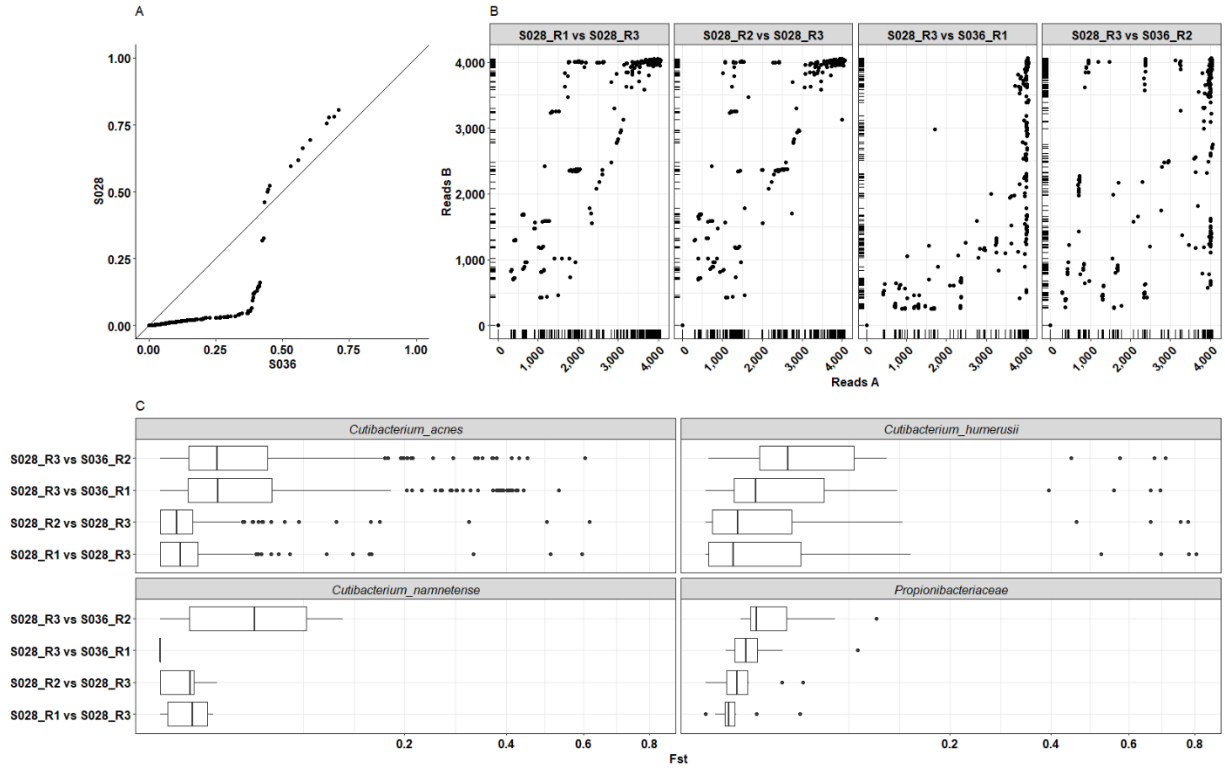


Figure 3. Comparisons of S028_R3 that were incorrectly classified as S036. A) A quantile-quantile plot of F_{ST} estimates for sample S028_R3 compared to individual S036. The distribution

of F_{ST} estimates between S028 (y-axis) and S036 (x-axis) and from comparing S028_R3 to other technical replicates. The F_{ST} estimates were computed for SNPs that were orthologous in at least two samples. The main diagonal represents S028 and S036 having equal values of F_{ST} estimates. Points below the main diagonal represent a greater differentiation between S036 and S028, while points above the diagonal show greater differentiation within S028. B) Shows the first sample in the graph labeled on the x-axis and the second sample on the y-axis with the number of reads plotted for the SNPs. The ticks on the x and y-axis show the density of the corresponding area on the graph to provide clarity about the density of plotted points. Overall, S028_R3 had less read coverage for SNPs in common with S036 than with S028. C) A boxplot of the F_{ST} estimates for each pairwise comparison. The distribution of F_{ST} estimates for the 36 markers S036 and S028 had in common tend to have higher F_{ST} for sample comparisons within S028 than between S036.

The test data also had samples that were incorrectly classified by all three methods. Specifically, the samples from S014 and S042 were often classified as S044, S046, or to each other. While some of the highest-ranking F_{ST} estimates are higher between S014 and S042, overall, there were more SNPs with high F_{ST} estimates within the individual than between individuals. For all replicates of S014, there did not tend to be any notable differences in the reads between selected SNPs. For S042_R2 and R3, the incorrect classifications may be due to S042_R3 having low read coverage for selected SNPs.

Discussion

This study investigated the potential of selecting high F_{ST} markers to improve HID using the skin microbiome. previous work with the hidSkinPlex using presence/absence or nucleotide diversity with nearest neighbor or normalized logistic regression achieved accuracy rates between 54.20% – 100.00% when classifying eight individuals with samples from three body sites collected in triplicate (4). Woerner et al. (11) expanded the number of individuals to 51 and sampled from the non-dominant hand in triplicate; using the same panel they achieved accuracies of 78.00% and 83.70% using phylogenetic distance or nucleotide diversity, respectively, for classification with nearest neighbor machine learning approaches. There was a decrease in classification accuracy

classification when the sample size was increased to 51 individuals compared to the eight samples in Schmedes et al. (4). The study herein re-analyzed the same sequence data as in Woerner et al. (11) with a novel method for SNP selection based on F_{ST} estimates and SVM and achieved higher accuracies (p-value = 0.03, chi-squared test, comparing the most accurate approaches in both studies). The accuracies of the three F_{ST} SNP selection methods also have increased accuracies compared to any of the previous studies using the targeted hidSkinPlex sequence data.

Three methods for selecting the highest-ranking F_{ST} estimations were used to assess how SNPs from the skin microbiome may be chosen for HID. All three methods of selecting informative SNPs had high classification accuracies. The *per marker* method achieved the highest accuracy (94.70%) which indicates that inclusion of more taxa potentially could increase classification accuracies. The *per marker* method allowed for the broadest selection of markers and SNPs in common in each single pairwise comparison, resulting in the method's higher accuracy. The *overall* method performed well with a 92.00% accuracy even though the number of SNPs used for analysis was less than the *per marker* method. While the *selected* method had the lowest accuracy of 88.00%, even though it initially had the highest training accuracy at 98.70%, the method still showed that a predetermined panel of chosen SNPs potentially could include or exclude a particular individual as the donor of a sample. An additional increase in accuracy might be achieved if minimum requirements were implemented to remove poor-quality samples. The results of this study provide support that using high-ranking F_{ST} estimates to select SNPs with SVM increased accuracies of classification to 94.70% and can potentially be used in a similar fashion as AIM are in human populations analyses.

The investigation into S028_R3 in the training data and S042_R3 in the test data set suggested that low read coverage and low diversity of a sample might impact classification

accuracy. If one of the three replicates from an individual has low read coverage and/or low diversity, the ability to correctly classify other replicates from the same individual may be impacted. Perhaps implementing minimum thresholds for analyzing a sample may eliminate poor quality samples from being searched. Additional research on potential minimum requirements, such as overall read coverage and depth and the number of total SNPs, may reduce the number of false positives (or for now better stated as adventitious hits). For the test data set, individual S014 and S042 were incorrectly classified in all three methods of SNP selection. Individuals S014 and S042 were also incorrectly classified to some degree by Woerner et al. (29) for both classification methods tested in their study. This observation suggests that replicates of individuals S014 and S042 may have been switched, contamination may have occurred during handling or processing, and/or that these individuals share a genetically and taxonomically similar microbiome. It is also possible that the SNPs selected for distinction individuals still need refinement and/or that thresholds for minimum data requirements need to be considered further. Additionally, studies need to be performed to determine why a few high F_{ST} SNPs could impact incorrect classification when the data as a whole support the correct classification.

Although the performance decreased with the test set, the *selected* method is of particular interest in that it provides a pre-determined set of SNPs to be used in every classification of the unknown samples. For the optimized parameter of 150 SNPs there were only two species and one family level marker represented, which were *Cutibacterium acnes*, *Cutibacterium humerusii*, and *Propionibacteriaceae.* These two species and one family level marker are common and abundant on the human skin and often have multiple subspecies or strains within individuals (28). The decrease in accuracy from 98.70% in the training data to 88.00% in the test data is most likely due to overfitting, both in the SNP ascertainment and in the SVM model itself. With more data for

training, it may be possible to adjust the pre-determined SNPs, but some level of overfitting will likely persist. A pre-determined panel would allow for the redesign of the hidSkinPlex to reduce the number and size of the markers in the panel with a potential increase in assay robustness.

Using F_{ST} estimates permitted selection of SNPs to be input into an SVM model. With a refined MPS targeted skin microbiome panel it will also be possible to further investigate how the SNPs of specific microorganisms change due to environment, health-status, and other external factors. Additionally, refinement of informative SNPs may provide an increase in the accuracy to include or exclude an individual as a potential contributor of a microbiome sample. The human skin microbiome has the potential to be supportive evidence to more traditional DNA evidence for law enforcement.

MATERIAL AND METHODS

Samples. Targeted sequence data from samples originally described in Woerner et al. (29) were used in this study. Briefly, skin swabs from 51 individuals were collected in triplicate from the non-dominant hand (Hp) of each individual (n = 153, replicates R1, R2, and R3). These samples were then analyzed using the hidSkinPlex, a targeted genome sequencing panel (4) drawn from the MetaPhlAn2 database (34). This panel targets 22 clades, with genus to subspecies level information, comprising 286 markers that were determined to be abundant and relatively stable on the human skin (35). The University of North Texas Health Science Center Institutional Review Board approved the collection and analyses of these samples.

Sequence data and analysis. All fastq files from the MiSeq were trimmed with cutadapt (36) to remove bases with a quality score less than 20 and reads less than 50 bases long as described in Woerner et al. (29). MetaPhlAn2 (34) was used to align sequence reads to the MetaPhlAn2

reference database. Samtools (37) was used to calculate read depth and coverage, and generate base pileups for each aligned marker in the hidSkinPlex panel.

Computation, Statistics and F_{ST} estimation. All statistics were performed in the R (v. 3.4.2) (38) or Python programming languages (v. 2.7.17, Python Software Foundation, https://www.python.org/) with plots created by ggplot2 (39). Welch two-sample *t*-tests and McNemar's chi-squared test were performed using the *stats* package (38). Hudson et al. (40) proposed estimating F_{ST} as $F_{ST} = 1-(H_w/H_b)$, where H_w is the mean number of pairwise differences within a population, and H_b is the mean number of pairwise differences between two populations (40). F_{ST} was estimated for all relevant nucleotide positions with a read depth minimum of one, it is worth noting that F_{ST} is only defined when $H_b > 0$ and that a minimum read depth parameter was optimized in the machine learning approach. When estimating F_{ST} the two samples (i.e., two populations) must each have at least one orthologous SNP being compared and have >1x read depth for the analysis (for example sample A at SNP position 25 has 2 read of A, and sample B has 2 read of C) an additional read depth parameter was optimized during the analysis of the training data set. Then a three-fold cross validation holding out one of the technical replicates was performed.

Machine learning strategy. A training set was used to optimize the linear support vector machine (SVM) C hyper-parameter, as well as a threshold on a maximum number of SNPs and minimum read depth. The test data set is used to determine how the SVM performed on unseen data. The training data set comprises 26 samples in triplicate (S001 – S012 and S025 – S037, where S0## represents an individual), and the test data set consists of 25 samples (S013 – S024 and S038 – S051).

The SVM approach embeds the distance (F_{ST}) between two individuals relative to a single query point into the Cartesian coordinate system. The embedding begins by considering four samples, two samples for each class (a class represents two samples from the same individual) and selecting the highest-ranking SNPs for each sample compared to the "unknown" sample. While embedding distances in the Cartesian coordinate system is not in general possible without error or loss, it is possible to use distance with a binary classifier when the distance is constrained to a single (query) point. A further benefit of the approach is it can be trained only on comparable data, in this case SNPs, between just the two samples and the query, in contrast to requiring the presence of each SNP in all samples. This allows the SVM to handle dropout in a way that avoids imputation and uses the variants to separate two individuals based on their common microbes.

Each comparison between two samples (one of them being the unknown data point), selected the highest-ranking F_{ST} estimates (i.e., SNPs) based on the selection method (i.e., o*verall, per marker, or selected*). After SNP selection, a matrix with the four samples (rows) and the selected SNPs (columns) was formed. If any SNP was not present in the other (up to three) comparisons, because it was not present as a high-ranking SNP, it was filled in with the F_{ST} estimate from the original data that met the minimum read requirement. Missing data was filled in with zero. F_{ST} values for common markers for all four comparisons were input into an in-house SVM code that used LibSVM (v. 1.7-3) (41) (R package e1071) as a feature vector with two labeled classes and a single unlabeled sample. The unknown sample was then provided as a vector of zeros as an additional feature vector to represent F_{ST} estimates of the unknown sample when compared to itself. The SVM provided a prediction about which of the two potential classes the unknown sample belonged and provided a percentage representing the SVM's confidence in its prediction. Each time the SVM made a classification to a particular class, the class was given a

vote of one. The total votes were tallied at the end. The total number of votes for each class was then used to rank the classes of the unknown sample.

Three approaches to select SNPs for analysis were developed to determine which method would provide the highest accuracy. Each method of high F_{ST} selection focused on a distinct approach to provide insight into whether the number of highest-ranking F_{ST} estimates increases classification accuracy or a common set of markers would more effectively improve accuracy of unknown sample prediction. The first approach, *overall*, selected either up to 500, 1,000, or 2,000 SNPs with the highest-ranked F_{ST} estimates across all markers, but not from any specific marker, to determine the minimum number of SNPs that could be used and still provide accurate classification. The second method, F_{ST} *per marker*, selected either 5, 10, or 25 SNPs contained within a marker with the highest-ranking F_{ST} per each marker common between the two populations that were compared. The third method, called *selected* F_{ST}, used all F_{ST} estimates with reads greater than 10 from the training samples to select SNPs that were seen most often in pairwise comparisons and had the highest-ranking F_{ST} estimates. The number of SNPs selected with the highest-ranking F_{ST} estimates were set at 50 to 2,000. The *selected* method chose SNPs by arranging F_{ST} estimates in descending order for each marker seen in all pairwise comparison in the training data set. All three selection methods were optimized under the objective of maximizing classification accuracy.

Parameter optimization. Three parameters were varied for all SVM models. The three parameters were the number of SNPs with the highest-ranking F_{ST} estimates in a pairwise comparison, the minimum reads at each SNP compared, and the cost (C-parameter) for the linear SVM. The number of SNPs selected with the highest-ranking F_{ST} estimates depended on which method was used, i.e., F_{ST} *per marker, overall* $F_{ST,}$ and *selected* $F_{ST.}$ A minimum read depth

threshold was assessed with each approach, and the thresholds were: 10, 100, 250, 500, or 1,000. Cost, the degree of misclassification allowed in the SVM, was set at 0.1, 1, 10, 100, or 1,000. The selection of optimal parameters for each F_{ST} selection method was evaluated by looking at the number of times each possible combination of all three parameters was used to predict 78 unknown samples with SVM. The accuracy was determined by the number of times that the unknown sample was predicted correctly (i.e., the highest rank).

DATA AVAILABILITY

Custom R and Python scripts can be accessed at https://github.com/CardiShire/PopulationInformativeMarkers.

ACKNOWLEDGEMENTS

We thank Sarah Schmedes for the design of the hidSkinPlex and the initial development of sample processing. Additionally, Angie Ambers, Rachel Kieser, Frank Wendt, Nicole Novroski, and Jonathan King for the countless hours they contributed to collecting/processing samples and providing feedback on the next steps for HID using the skin microbiome. Last but not least, we thank Utpal Smart, Sammed Mandape, Ben Crysup, and Jonathan King for all the time they spent advising on code and debugging support.

This study was supported in part by the National Institute of Justice, Award Number 2015-NE-BX-K006 and 2020-R2-CX-0046.

BIBLIOGRAPHY

1. Sender R, Fuchs S, Milo R. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol 14:e1002533.
2. Sender R, Fuchs S, Milo R. 2016. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell 164:337-40.
3. Grice EA, Kong HH, Renaud G, Young AC, Program NCS, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre JA. 2008. A diversity profile of the human skin microbiota. Genome Res 18:1043-50.
4. Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. 2018. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Sci Int Genet 32:50-61.
5. Hampton-Marcell JT, Larsen P, Anton T, Cralle L, Sangwan N, Lax S, Gottel N, Salas-Garcia M, Young C, Duncan G, Lopez JV, Gilbert JA. 2020. Detecting personal microbiota signatures at artificial crime scenes. Forensic Sci Int 313:110351.
6. Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen JA, Gilbert JA. 2015. Forensic analysis of the microbiome of phones and shoes. Microbiome 3:21.
7. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S, Metcalf JL, Ursell LK, Vazquez-Baeza Y, Van Treuren W, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, Gilbert JA. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science 345:1048-52.
8. Lax S NC, Gilbert JA. 2015. Our interface with the built environment: immunity and the indoor microbiota. Trends Immunol.

9. Richardson M, Gottel N, Gilbert JA, Lax S. 2019. Microbial Similarity between Students in a Common Dormitory Environment Reveals the Forensic Potential of Individual Microbial Signatures. mBio 10:e01054-19.

10. Luongo JC, Barberán A, Hacker-Cary R, Morgan EE, Miller SL, Fierer N. 2017. Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. Indoor Air 27:338-344.

11. Adams RI, Bateman AC, Bik HM, Meadow JF. 2015. Microbiota of the indoor environment: a meta-analysis. Microbiome 3:49.

12. Fujiyoshi S, Tanaka D, Maruyama F. 2017. Transmission of Airborne Bacteria across Built Environments and Its Measurement Standards: A Review. Front Microbiol 8:2336.

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

14. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, Brown GZ, Green JL, Bohannan BJ. 2014. Bacterial communities on classroom surfaces vary with human contact. Microbiome 2:7.

15. 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 107:6477-81.

16. Park J, Kim SJ, Lee J-A, Kim JW, Kim SB. 2017. Microbial forensic analysis of human-associated bacteria inhabiting hand surface. Forensic Science International: Genetics Supplement Series 6:e510-e512.

17. Watanabe H, Nakamura I, Mizutani S, Kurokawa Y, Mori H, Kurokawa K, Yamada T. 2018. Minor taxa in human skin microbiome contribute to the personal identification. PLoS One 13:e0199947.

18. Yang J, Tsukimi T, Yoshikawa M, Suzuki K, Takeda T, Tomita M, Fukuda S. 2019. Cutibacterium acnes (Propionibacterium acnes) 16S rRNA Genotyping of Microbial Samples from Possessions Contributes to Owner Identification. mSystems 4:e00594-19.

19. Doleckova I, Capova A, Machkova L, Moravcikova S, Maresova M, Velebny V. 2020. Seasonal variations in the skin parameters of Caucasian women from Central Europe. Skin Res Technol n/a.

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

21. Song SJ, Lauber C, Costello EK, Lozupone CA, 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.

22. Neckovic A, van Oorschot RAH, Szkuta B, Durdle A. 2020. Investigation of direct and indirect transfer of microbiomes between individuals. Forensic Sci Int Genet 45:102212.

23. Bosshard PP, Zbinden R, Abels S, Boddinghaus B, Altwegg M, Bottger EC. 2006. 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. J Clin Microbiol 44:1359-66.

24. Mignard S, Flandrois JP. 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. J Microbiol Methods 67:574-81.

25. Fox GE, Wisotzkey JD, Jurtshuk P, Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166-70.

26. Lee S-Y, Woo S-K, Lee S-M, Eom Y-B. 2016. Forensic analysis using microbial community between skin bacteria and fabrics. Toxicology and Environmental Health Sciences 8:263-270.

27. Gu Y, Zha L, Yun L. 2017. Potential usefulness of SNP in the 16S rRNA gene serving as informative microbial marker for forensic attribution. Forensic Science International: Genetics Supplement Series 6:e451-e452.

28. Oh J, Byrd AL, Park M, Program NCS, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854-66.

29. Woerner AE, Novroski NMM, Wendt FR, Ambers A, Wiley R, Schmedes SE, Budowle B. 2019. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci Int Genet 38:130-139.

30. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR. 2014. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 10:23-32.

31. Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Borsting C, Johansen P, Fondevila M, Morling N, Schneider P, Consortium EU-N, Carracedo A, Lareu MV. 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic Sci Int Genet 11:13-25.

32. Wright S. 1949. The Genetical Structure of Populations. Annals of Eugenics 15:323-354.

33. Fay MP. 2010. Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. R Journal 2:53-58.

34. 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 12:902-3.

35. Schmedes SE, Woerner AE, Budowle B. 2017. Forensic Human Identification Using Skin Microbiomes. Appl Environ Microbiol 83.

36. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:3.

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.

38. Team RC. 2013. R: A Language and Environment for Statistical Computing, *on* R Foundation for Statistical Computing. http://www.R-project.org/. Accessed 2/.

39. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K. 2016. ggplot2: Elegant Graphics for Data Analysis, vol 2018. Springer-Verlag New York.

40. Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-9.

41. Meyer D, Evgenia Dimitriadou, Hornik K, Weingessel A, Leisch F. 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3. https://CRAN.R-project.org/package=e1071.

SUPPLEMENTAL

Table S1. The range of nucleotides and variants seen in each sample comparison for the training data set. F_{ST} estimations were calculated for each nucleotide in common between two samples (Individual A and Individual B). Then all F_{ST} values greater than 0 ($F_{ST} > 0$) were used to calculate the minimum, median, mean, and maximum. Additionally, the number of F_{ST} estimates higher than 0.10, 0.25, and 0.50 were determined for each comparison. S0## represents an individual, and R# represents the replicate.

Marker	SNP Position	Species	Marker Length
gi|295129529|ref|NC_014039.1|:c1439020-1438442	111	Cutibacterium_acnes	579
gi|295129529|ref|NC_014039.1|:c1439020-1438442	120	Cutibacterium_acnes	579
gi|295129529|ref|NC_014039.1|:c1439020-1438442	149	Cutibacterium_acnes	579
gi|295129529|ref|NC_014039.1|:c1439020-1438442	162	Cutibacterium_acnes	579
gi|295129529|ref|NC_014039.1|:c1439020-1438442	169	Cutibacterium_acnes	579
gi|295129529|ref|NC_014039.1|:c1439020-1438442	369	Cutibacterium_acnes	579
gi|365961730|ref|NC_016511.1|:2485446-2486162	325	Cutibacterium_acnes	717
gi|387502364|ref|NC_017535.1|:c1339878-1339075	597	Cutibacterium_acnes	804
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	103	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	113	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	114	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	120	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	121	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	71	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	98	Cutibacterium_humerusii	126
gi|395203690|ref|NZ_AFAM01000005.1|:c52756-52631	99	Cutibacterium_humerusii	126
gi|335050601|ref|NZ_AFIK01000014.1|:3050-3691	449	Cutibacterium_acnes	642
gi|335050601|ref|NZ_AFIK01000014.1|:315-1133	318	Cutibacterium_acnes	819
gi|335050601|ref|NZ_AFIK01000014.1|:315-1133	517	Cutibacterium_acnes	819
gi|335050796|ref|NZ_AFIK01000023.1|:c3954-3715	58	Cutibacterium_acnes	240
gi|335051382|ref|NZ_AFIK01000053.1|:c36245-34977	462	Cutibacterium_acnes	1269
gi|335051382|ref|NZ_AFIK01000053.1|:c36245-34977	595	Cutibacterium_acnes	1269
gi|335051382|ref|NZ_AFIK01000053.1|:c36245-34977	657	Cutibacterium_acnes	1269
gi|335051798|ref|NZ_AFIK01000065.1|:c4330-4001	192	Cutibacterium_acnes	330
gi|335052272|ref|NZ_AFIK01000082.1|:c111360-110575	222	Cutibacterium_acnes	786
gi|335052272|ref|NZ_AFIK01000082.1|:c111360-110575	390	Cutibacterium_acnes	786
gi|335052272|ref|NZ_AFIK01000082.1|:c111360-110575	610	Cutibacterium_acnes	786
gi|335052272|ref|NZ_AFIK01000082.1|:c111360-110575	700	Cutibacterium_acnes	786
gi|335053104|ref|NZ_AFIL01000010.1|:c43071-42837	154	Cutibacterium_acnes	235
gi|335053104|ref|NZ_AFIL01000010.1|:c43071-42837	209	Cutibacterium_acnes	235
gi|335053104|ref|NZ_AFIL01000010.1|:c43071-42837	56	Cutibacterium_acnes	235
gi|335053104|ref|NZ_AFIL01000010.1|:c43071-42837	79	Cutibacterium_acnes	235
gi|335053207|ref|NZ_AFIL01000016.1|:c75436-75296	141	Cutibacterium_acnes	141
gi|335053685|ref|NZ_AFIL01000030.1|:c58004-57372	146	Cutibacterium_acnes	633

gi|335053685|ref|NZ_AFIL01000030.1|:c58004-57372	307	Cutibacterium_acnes	633
gi|335054110|ref|NZ_AFIL01000040.1|:4048-4263	133	Cutibacterium_acnes	216
gi|335054110|ref|NZ_AFIL01000040.1|:4048-4263	89	Cutibacterium_acnes	216
gi|335054139|ref|NZ_AFIL01000041.1|:c77880-77749	104	Cutibacterium_acnes	132
gi|335054139|ref|NZ_AFIL01000041.1|:c77880-77749	57	Cutibacterium_acnes	132
gi|335054139|ref|NZ_AFIL01000041.1|:c77880-77749	74	Cutibacterium_acnes	132
gi|335054309|ref|NZ_AFIL01000044.1|:65842-65994	56	Cutibacterium_acnes	153
gi|335054520|ref|NZ_AFIL01000051.1|:c25042-24929	71	Cutibacterium_acnes	114
gi|335055047|ref|NZ_AFIL01000069.1|:c9632-8838	308	Cutibacterium_acnes	795
gi|335055047|ref|NZ_AFIL01000069.1|:c9632-8838	453	Cutibacterium_acnes	795
gi|335055047|ref|NZ_AFIL01000069.1|:c9632-8838	632	Cutibacterium_acnes	795
gi|335055061|ref|NZ_AFIL01000070.1|:3643-4386	252	Cutibacterium_acnes	744
gi|335055061|ref|NZ_AFIL01000070.1|:3643-4386	258	Cutibacterium_acnes	744
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	1042	Cutibacterium_acnes	1323
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	706	Cutibacterium_acnes	1323
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	824	Cutibacterium_acnes	1323
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	897	Cutibacterium_acnes	1323
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	924	Cutibacterium_acnes	1323
gi|342211239|ref|NZ_AFUK01000001.1|:665124-666446	978	Cutibacterium_acnes	1323
gi|342211239|ref|NZ_AFUK01000001.1|:c1255510-1255055	290	Cutibacterium_acnes	456
gi|342211239|ref|NZ_AFUK01000001.1|:c1376325-1376110	101	Cutibacterium_acnes	216
gi|342211239|ref|NZ_AFUK01000001.1|:c1715790-1715233	202	Cutibacterium_acnes	558
gi|342211239|ref|NZ_AFUK01000001.1|:c1715790-1715233	311	Cutibacterium_acnes	558
gi|342211239|ref|NZ_AFUK01000001.1|:c1845075-1844710	160	Cutibacterium_acnes	366
gi|342211239|ref|NZ_AFUK01000001.1|:c1936798-1936352	249	Cutibacterium_acnes	447
gi|552879811|ref|NZ_AXME01000001.1|:1088727-1089377	504	Cutibacterium_acnes	651
gi|552879811|ref|NZ_AXME01000001.1|:1146402-1146932	165	Cutibacterium_acnes	531
gi|552879811|ref|NZ_AXME01000001.1|:1286960-1287442	157	Cutibacterium_acnes	483
gi|552879811|ref|NZ_AXME01000001.1|:1286960-1287442	204	Cutibacterium_acnes	483
gi|552879811|ref|NZ_AXME01000001.1|:1286960-1287442	66	Cutibacterium_acnes	483
gi|552879811|ref|NZ_AXME01000001.1|:1431752-1431913	59	Cutibacterium_acnes	162
gi|552879811|ref|NZ_AXME01000001.1|:1431752-1431913	61	Cutibacterium_acnes	162
gi|552879811|ref|NZ_AXME01000001.1|:1431752-1431913	92	Cutibacterium_acnes	162
gi|552879811|ref|NZ_AXME01000001.1|:40840-41742	476	Cutibacterium_acnes	903
gi|552879811|ref|NZ_AXME01000001.1|:49241-49654	219	Cutibacterium_acnes	414
gi|552879811|ref|NZ_AXME01000001.1|:49241-49654	285	Cutibacterium_acnes	414
gi|552879811|ref|NZ_AXME01000001.1|:49241-49654	303	Cutibacterium_acnes	414
gi|552879811|ref|NZ_AXME01000001.1|:587256-587825	301	Cutibacterium_acnes	570
gi|552879811|ref|NZ_AXME01000001.1|:587256-587825	32	Cutibacterium_acnes	570
gi|552879811|ref|NZ_AXME01000001.1|:587256-587825	375	Cutibacterium_acnes	570
gi|552879811|ref|NZ_AXME01000001.1|:587256-587825	507	Cutibacterium_acnes	570
gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	103	Cutibacterium_acnes	207
gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	57	Cutibacterium_acnes	207

gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	58	Cutibacterium_acnes	207
gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	67	Cutibacterium_acnes	207
gi|552879811|ref|NZ_AXME01000001.1|:655649-655855	77	Cutibacterium_acnes	207
gi|552879811|ref|NZ_AXME01000001.1|:865400-865597	58	Cutibacterium_acnes	198
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	147	Cutibacterium_acnes	270
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	149	Cutibacterium_acnes	270
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	197	Cutibacterium_acnes	270
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	200	Cutibacterium_acnes	270
gi|552879811|ref|NZ_AXME01000001.1|:990664-990933	89	Cutibacterium_acnes	270
gi|552879811|ref|NZ_AXME01000001.1|:c1552174-1551533	256	Cutibacterium_acnes	642
gi|552879811|ref|NZ_AXME01000001.1|:c1599141-1598893	66	Cutibacterium_acnes	249
gi|552879811|ref|NZ_AXME01000001.1|:c1599141-1598893	69	Cutibacterium_acnes	249
gi|552879811|ref|NZ_AXME01000001.1|:c1599141-1598893	93	Cutibacterium_acnes	249
gi|552879811|ref|NZ_AXME01000001.1|:c1820429-1820292	53	Cutibacterium_acnes	138
gi|552879811|ref|NZ_AXME01000001.1|:c1820429-1820292	67	Cutibacterium_acnes	138
gi|552879811|ref|NZ_AXME01000001.1|:c1820429-1820292	69	Cutibacterium_acnes	138
gi|552879811|ref|NZ_AXME01000001.1|:c1820429-1820292	76	Cutibacterium_acnes	138
gi|552879811|ref|NZ_AXME01000001.1|:c2014536-2014075	411	Cutibacterium_acnes	462
gi|552879811|ref|NZ_AXME01000001.1|:c2014536-2014075	414	Cutibacterium_acnes	462
gi|552879811|ref|NZ_AXME01000001.1|:c2014536-2014075	417	Cutibacterium_acnes	462
gi|552879811|ref|NZ_AXME01000001.1|:c2014536-2014075	423	Cutibacterium_acnes	462
gi|552879811|ref|NZ_AXME01000001.1|:c2447430-2446870	136	Cutibacterium_acnes	561
gi|552879811|ref|NZ_AXME01000001.1|:c2447430-2446870	396	Cutibacterium_acnes	561
gi|552879811|ref|NZ_AXME01000001.1|:c2447430-2446870	495	Cutibacterium_acnes	561
gi|552879811|ref|NZ_AXME01000001.1|:c31864-31571	165	Cutibacterium_acnes	294
gi|552879811|ref|NZ_AXME01000001.1|:c31864-31571	98	Cutibacterium_acnes	294
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	144	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	168	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	172	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	336	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	357	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	369	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	409	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	416	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	438	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	474	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	90	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	93	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1231251-1231871	99	Cutibacterium_acnes	621
gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	42	Cutibacterium_acnes	252
gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	56	Cutibacterium_acnes	252
gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	68	Cutibacterium_acnes	252
gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	79	Cutibacterium_acnes	252

gi|552891898|ref|NZ_AXMG01000001.1|:1440218-1440469	80	Cutibacterium_acnes	252
gi|552891898|ref|NZ_AXMG01000001.1|:793445-793843	261	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:834824-835255	176	Cutibacterium_acnes	432
gi|552891898|ref|NZ_AXMG01000001.1|:834824-835255	378	Cutibacterium_acnes	432
gi|552891898|ref|NZ_AXMG01000001.1|:99114-99290	107	Cutibacterium_acnes	177
gi|552891898|ref|NZ_AXMG01000001.1|:c1328090-1327596	314	Cutibacterium_acnes	495
gi|552891898|ref|NZ_AXMG01000001.1|:c1328090-1327596	426	Cutibacterium_acnes	495
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	144	Cutibacterium_acnes	603
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	219	Cutibacterium_acnes	603
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	327	Cutibacterium_acnes	603
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	333	Cutibacterium_acnes	603
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	339	Cutibacterium_acnes	603
gi|552891898|ref|NZ_AXMG01000001.1|:c1443707-1443105	495	Cutibacterium_acnes	603
gi|552891898|ref|NZ_AXMG01000001.1|:c1945194-1944973	111	Cutibacterium_acnes	222
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	120	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	126	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	162	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	166	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	185	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	237	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	318	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	388	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	389	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2126720-2126193	390	Cutibacterium_acnes	528
gi|552891898|ref|NZ_AXMG01000001.1|:c2312839-2311925	221	Cutibacterium_acnes	915
gi|552891898|ref|NZ_AXMG01000001.1|:c2312839-2311925	281	Cutibacterium_acnes	915
gi|552891898|ref|NZ_AXMG01000001.1|:c2312839-2311925	662	Cutibacterium_acnes	915
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	120	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	135	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	140	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	177	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	201	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	227	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	234	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	31	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	354	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2382295-2381897	54	Cutibacterium_acnes	399
gi|552891898|ref|NZ_AXMG01000001.1|:c2429318-2428110	657	Cutibacterium_acnes	1209
gi|552891898|ref|NZ_AXMG01000001.1|:c2429318-2428110	696	Cutibacterium_acnes	1209
gi|552891898|ref|NZ_AXMG01000001.1|:c2429318-2428110	702	Cutibacterium_acnes	1209
gi|552895565|ref|NZ_AXMI01000001.1|:619555-620031	242	Cutibacterium_acnes	477
gi|552895565|ref|NZ_AXMI01000001.1|:c14352-13837	235	Cutibacterium_acnes	516
gi|552895565|ref|NZ_AXMI01000001.1|:c29469-28930	195	Cutibacterium_acnes	540

gi|552895565|ref|NZ_AXMI01000001.1|:c29469-28930	390	Cutibacterium_acnes	540
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	127	Cutibacterium_acnes	1116
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	135	Cutibacterium_acnes	1116
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	144	Cutibacterium_acnes	1116
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	271	Cutibacterium_acnes	1116
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	456	Cutibacterium_acnes	1116
gi|552895565|ref|NZ_AXMI01000001.1|:c443438-442323	468	Cutibacterium_acnes	1116
gi|552896371|ref|NZ_AXMI01000002.1|:319095-319601	382	Cutibacterium_acnes	507
gi|552896371|ref|NZ_AXMI01000002.1|:525312-525770	175	Cutibacterium_acnes	459
gi|552896371|ref|NZ_AXMI01000002.1|:525312-525770	356	Cutibacterium_acnes	459
gi|552896371|ref|NZ_AXMI01000002.1|:525312-525770	363	Cutibacterium_acnes	459
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	147	Cutibacterium_acnes	600
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	219	Cutibacterium_acnes	600
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	381	Cutibacterium_acnes	600
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	448	Cutibacterium_acnes	600
gi|552896371|ref|NZ_AXMI01000002.1|:674988-675587	483	Cutibacterium_acnes	600
gi|552896371|ref|NZ_AXMI01000002.1|:721564-722400	543	Cutibacterium_acnes	837
gi|552896371|ref|NZ_AXMI01000002.1|:837080-837400	114	Cutibacterium_acnes	321
gi|552896371|ref|NZ_AXMI01000002.1|:837080-837400	49	Cutibacterium_acnes	321
gi|552896371|ref|NZ_AXMI01000002.1|:c671938-670697	220	Cutibacterium_acnes	1242
gi|552896688|ref|NZ_AXMI01000003.1|:232201-232740	172	Cutibacterium_acnes	540
gi|552896688|ref|NZ_AXMI01000003.1|:232201-232740	282	Cutibacterium_acnes	540
gi|552896688|ref|NZ_AXMI01000003.1|:232201-232740	507	Cutibacterium_acnes	540
gi|552896688|ref|NZ_AXMI01000003.1|:232201-232740	76	Cutibacterium_acnes	540
gi|552897201|ref|NZ_AXMI01000004.1|:13568-14401	336	Cutibacterium_acnes	834
gi|552897201|ref|NZ_AXMI01000004.1|:13568-14401	519	Cutibacterium_acnes	834
gi|552897201|ref|NZ_AXMI01000004.1|:13568-14401	657	Cutibacterium_acnes	834
gi|552897201|ref|NZ_AXMI01000004.1|:48085-48816	426	Cutibacterium_acnes	732
gi|552897201|ref|NZ_AXMI01000004.1|:48085-48816	453	Cutibacterium_acnes	732
gi|552897201|ref|NZ_AXMI01000004.1|:48085-48816	512	Cutibacterium_acnes	732
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	172	Cutibacterium_acnes	813
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	228	Cutibacterium_acnes	813
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	300	Cutibacterium_acnes	813
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	481	Cutibacterium_acnes	813
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	486	Cutibacterium_acnes	813
gi|552897201|ref|NZ_AXMI01000004.1|:c102788-101976	670	Cutibacterium_acnes	813
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	118	Cutibacterium_acnes	555
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	210	Cutibacterium_acnes	555
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	219	Cutibacterium_acnes	555
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	474	Cutibacterium_acnes	555
gi|552897201|ref|NZ_AXMI01000004.1|:c231437-230883	54	Cutibacterium_acnes	555
gi|552897201|ref|NZ_AXMI01000004.1|:c577292-575922	784	Cutibacterium_acnes	1371
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	107	Cutibacterium_acnes	447

gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	177	Cutibacterium_acnes	447
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	284	Cutibacterium_acnes	447
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	286	Cutibacterium_acnes	447
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	337	Cutibacterium_acnes	447
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	338	Cutibacterium_acnes	447
gi|552902020|ref|NZ_AXMK01000001.1|:c1228696-1228250	89	Cutibacterium_acnes	447
gi|422552858|ref|NZ_GL383469.1|:c216727-215501	1006	Cutibacterium_acnes	1227
gi|422552858|ref|NZ_GL383469.1|:c216727-215501	777	Cutibacterium_acnes	1227
gi|422552858|ref|NZ_GL383469.1|:c216727-215501	788	Cutibacterium_acnes	1227
gi|422482616|ref|NZ_GL383714.1|:170052-170369	99	Cutibacterium_acnes	318
gi|422500804|ref|NZ_GL383759.1|:c166532-166311	43	Cutibacterium_acnes	222
gi|422500804|ref|NZ_GL383759.1|:c166532-166311	79	Cutibacterium_acnes	222
gi|422500804|ref|NZ_GL383759.1|:c166532-166311	86	Cutibacterium_acnes	222
gi|422500804|ref|NZ_GL383759.1|:c166532-166311	87	Cutibacterium_acnes	222
gi|422496709|ref|NZ_GL383802.1|:56803-56916	35	Cutibacterium_acnes	114
gi|422499020|ref|NZ_GL383811.1|:10443-11039	105	Cutibacterium_acnes	597
gi|422499020|ref|NZ_GL383811.1|:10443-11039	255	Cutibacterium_acnes	597
gi|422499020|ref|NZ_GL383811.1|:10443-11039	294	Cutibacterium_acnes	597
gi|422499020|ref|NZ_GL383811.1|:10443-11039	408	Cutibacterium_acnes	597
gi|422512600|ref|NZ_GL383846.1|:26161-26922	179	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	202	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	222	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	228	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	331	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	385	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	411	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	485	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	487	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	522	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	536	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	565	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	566	Cutibacterium_acnes	762
gi|422512600|ref|NZ_GL383846.1|:26161-26922	584	Cutibacterium_acnes	762
gi|422423570|ref|NZ_GL384259.1|:c300859-299957	310	Cutibacterium_acnes	903
gi|422423570|ref|NZ_GL384259.1|:c300859-299957	371	Cutibacterium_acnes	903
gi|422423570|ref|NZ_GL384259.1|:c300859-299957	484	Cutibacterium_acnes	903
gi|422423570|ref|NZ_GL384259.1|:c300859-299957	498	Cutibacterium_acnes	903
gi|422436532|ref|NZ_GL384462.1|:c297812-297150	489	Cutibacterium_acnes	663
gi|422385765|ref|NZ_GL878448.1|:c80834-80607	183	Cutibacterium_acnes	228
gi|422385765|ref|NZ_GL878448.1|:c80834-80607	33	Cutibacterium_acnes	228
gi|422385765|ref|NZ_GL878448.1|:c80834-80607	45	Cutibacterium_acnes	228
gi|422386402|ref|NZ_GL878455.1|:c805995-805537	178	Cutibacterium_acnes	459
gi|422386402|ref|NZ_GL878455.1|:c805995-805537	435	Cutibacterium_acnes	459

gi|422386402|ref|NZ_GL878455.1|:c805995-805537	75	Cutibacterium_acnes	459
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	444	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	520	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	530	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	552	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	678	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	705	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	940	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1103467-1104744	948	Cutibacterium_acnes	1278
gi|355707189|ref|NZ_JH376566.1|:1105369-1105965	540	Cutibacterium_acnes	597
gi|355707189|ref|NZ_JH376566.1|:1105369-1105965	58	Cutibacterium_acnes	597
gi|355707189|ref|NZ_JH376566.1|:1105369-1105965	65	Cutibacterium_acnes	597
gi|355707189|ref|NZ_JH376566.1|:507019-507612	166	Cutibacterium_acnes	594
gi|355707384|ref|NZ_JH376567.1|:190789-191232	169	Cutibacterium_acnes	444
gi|355707384|ref|NZ_JH376567.1|:190789-191232	237	Cutibacterium_acnes	444
gi|355707384|ref|NZ_JH376567.1|:251291-251998	417	Cutibacterium_acnes	708
gi|355707384|ref|NZ_JH376567.1|:251291-251998	638	Cutibacterium_acnes	708
gi|355707384|ref|NZ_JH376567.1|:251291-251998	66	Cutibacterium_acnes	708
gi|355707384|ref|NZ_JH376567.1|:251291-251998	74	Cutibacterium_acnes	708
gi|355707384|ref|NZ_JH376567.1|:251291-251998	84	Cutibacterium_acnes	708
gi|355707384|ref|NZ_JH376567.1|:592116-592328	25	Cutibacterium_acnes	213
gi|355707384|ref|NZ_JH376567.1|:c388018-387605	138	Cutibacterium_acnes	414
gi|355707384|ref|NZ_JH376567.1|:c388018-387605	252	Cutibacterium_acnes	414
gi|355707384|ref|NZ_JH376567.1|:c388018-387605	68	Cutibacterium_acnes	414
gi|355707384|ref|NZ_JH376567.1|:c388018-387605	69	Cutibacterium_acnes	414
gi|355708280|ref|NZ_JH376568.1|:c255689-255105	356	Cutibacterium_acnes	585
gi|355708280|ref|NZ_JH376568.1|:c255689-255105	461	Cutibacterium_acnes	585
gi|355708440|ref|NZ_JH376569.1|:c80380-79448	375	Cutibacterium_acnes	933
gi|355708440|ref|NZ_JH376569.1|:c80380-79448	537	Cutibacterium_acnes	933
gi|355708440|ref|NZ_JH376569.1|:c80380-79448	575	Cutibacterium_acnes	933
gi|552875787|ref|NZ_KI515684.1|:459339-460115	103	Cutibacterium_acnes	777
gi|552875787|ref|NZ_KI515684.1|:459339-460115	315	Cutibacterium_acnes	777
gi|552875787|ref|NZ_KI515684.1|:c325537-325361	23	Cutibacterium_acnes	177
gi|552875787|ref|NZ_KI515684.1|:c44215-43715	187	Cutibacterium_acnes	501
gi|552875787|ref|NZ_KI515684.1|:c44215-43715	276	Cutibacterium_acnes	501
gi|552875787|ref|NZ_KI515684.1|:c44215-43715	358	Cutibacterium_acnes	501
gi|552875787|ref|NZ_KI515684.1|:c44215-43715	61	Cutibacterium_acnes	501
gi|552875787|ref|NZ_KI515684.1|:c488989-488798	105	Cutibacterium_acnes	192
gi|552875787|ref|NZ_KI515684.1|:c488989-488798	37	Cutibacterium_acnes	192
gi|552875787|ref|NZ_KI515684.1|:c488989-488798	93	Cutibacterium_acnes	192
gi|552875787|ref|NZ_KI515684.1|:c584270-583890	201	Cutibacterium_acnes	381
gi|552875787|ref|NZ_KI515684.1|:c96934-96368	348	Cutibacterium_acnes	567
gi|552875787|ref|NZ_KI515684.1|:c96934-96368	400	Cutibacterium_acnes	567

gi|552875787|ref|NZ_KI515684.1|:c96934-96368	517	Cutibacterium_acnes	567
gi|552875787|ref|NZ_KI515684.1|:c96934-96368	518	Cutibacterium_acnes	567
gi|552876418|ref|NZ_KI515685.1|:133418-133666	213	Cutibacterium_acnes	249
gi|552876418|ref|NZ_KI515685.1|:187493-188140	183	Cutibacterium_acnes	648
gi|552876418|ref|NZ_KI515685.1|:187493-188140	225	Cutibacterium_acnes	648
gi|552876418|ref|NZ_KI515685.1|:187493-188140	261	Cutibacterium_acnes	648
gi|552876418|ref|NZ_KI515685.1|:187493-188140	411	Cutibacterium_acnes	648
gi|552876418|ref|NZ_KI515685.1|:187493-188140	548	Cutibacterium_acnes	648
gi|552876418|ref|NZ_KI515685.1|:187493-188140	85	Cutibacterium_acnes	648
gi|552876418|ref|NZ_KI515685.1|:225601-226386	159	Cutibacterium_acnes	786
gi|552876418|ref|NZ_KI515685.1|:225601-226386	181	Cutibacterium_acnes	786
gi|552876418|ref|NZ_KI515685.1|:225601-226386	636	Cutibacterium_acnes	786
gi|552876418|ref|NZ_KI515685.1|:536580-547218	339	Cutibacterium_acnes	639
gi|552876418|ref|NZ_KI515685.1|:536580-547218	453	Cutibacterium_acnes	639
gi|552876418|ref|NZ_KI515685.1|:536580-547218	96	Cutibacterium_acnes	639
gi|552876418|ref|NZ_KI515685.1|:c743399-743001	100	Cutibacterium_acnes	399
gi|552876418|ref|NZ_KI515685.1|:c743399-743001	129	Cutibacterium_acnes	399
gi|552876418|ref|NZ_KI515685.1|:c743399-743001	144	Cutibacterium_acnes	399
gi|552876418|ref|NZ_KI515685.1|:c743399-743001	235	Cutibacterium_acnes	399
gi|552876418|ref|NZ_KI515685.1|:c849089-848304	167	Cutibacterium_acnes	786
gi|552876418|ref|NZ_KI515685.1|:c849089-848304	229	Cutibacterium_acnes	786
gi|552876418|ref|NZ_KI515685.1|:c849089-848304	660	Cutibacterium_acnes	786
gi|552876815|ref|NZ_KI515686.1|:323579-324514	188	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:323579-324514	312	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:323579-324514	453	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:323579-324514	528	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:323579-324514	549	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:323579-324514	596	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:323579-324514	771	Cutibacterium_acnes	936
gi|552876815|ref|NZ_KI515686.1|:613740-614315	105	Cutibacterium_acnes	576
gi|552876815|ref|NZ_KI515686.1|:c200743-199319	457	Cutibacterium_acnes	1425
gi|552876815|ref|NZ_KI515686.1|:c200743-199319	573	Cutibacterium_acnes	1425
gi|552876815|ref|NZ_KI515686.1|:c642879-642748	108	Cutibacterium_acnes	132
gi|552876815|ref|NZ_KI515686.1|:c642879-642748	94	Cutibacterium_acnes	132
gi|552904108|ref|NZ_KI518468.1|:464070-464315	92	Cutibacterium_acnes	246

Table S2. Votes and rank of classes for sample S028_R3 for *per marker* and *overall* method. The total number of votes possible via SVM for the training data set was 25 votes. Sample S028_R3 had no classification obtaining 25 votes. For classes that received the same number of votes, the mean prediction percentage was used to break the tie.

Votes	Classification	Rank
24	S036	1
23	S029	2
22	S007	3
22	S004	4
22	S025	5
20	S008	6
20	S006	7
18	S001	8
18	S010	9
16	S028	10

CHAPTER III

Determining informative microbial single nucleotide polymorphisms for human identification

Submitted to Applied and Environmental Microbiology

Allison J. Sherier
August E. Woerner
Bruce Budowle

ABSTRACT The skin microbiome is a highly abundant and relatively stable source of DNA that may be utilized for human identification (HID). In this study, a set of SNPs with a high mean estimated F_{ST} (> 0.1) and widespread abundance (found in ≥ 75% of samples compared) were selected from a diverse set of markers in the hidSkinPlex. The least absolute shrinkage and selection operator (LASSO) was used in a novel machine learning framework to generate a SNP panel and predict the human host from skin microbiome samples collected from the hand, manubrium, and foot. The framework was devised to emulate a new unknown person introduced to the algorithm and to match samples from that person against a population database. Unknown samples were classified with 96% accuracy (MCC = 0.954) in the test (n = 225 samples) data set. A final panel of informative SNPs was determined for HID (hidSkinPlex+) using all 51 individuals sampled at three body sites in triplicate. The hidSkinPlex+ is comprised of 365 SNPs and yielded prediction accuracy for the correct host of 95% (MCC = 0.949). The accuracy of the hidSkinPlex+ may be somewhat overestimated due to using 26 individuals from the training data set for the selection of the final panel. However, this accuracy still provides an indication of performance when tested on new samples.

IMPORTANCE One of the fundamental goals in forensic genetics is to identify the source of biological evidence. Methods for detecting human DNA have advanced and can be quite sensitive, but not all DNA samples are amenable to current methods. Yet, the human skin microbiome is a source of DNA with high copy numbers, and it has the potential for high discriminatory power. The hidSkinPlex has been used for HID; however, some aspects of the panel could be improved. Missing information is ambiguous as it is unclear if marker drop-out is a byproduct of a low-template sample or if the reasons for not observing a marker are biological. Such ambiguity may confound methods for HID, and, as such, an improved marker set (the hidSkinPlex+) was designed

that is considerably smaller and more robust to drop-out (365 SNPs contained in 135 markers) yet still can be used to accurately predict the human host.

KEYWORDS hidSkinPlex, skin microbiome, microbial forensics, human identification, massively parallel sequencing, machine learning, multinomial logistic regression, Wright's fixation index

INTRODUCTION

The human microbiome encompasses the fungi, bacteria, and viruses living on and in individuals and their surrounding environment. The interplay of genetics and environment results in each person having a skin microbiome that is suggested to be unique (1-4). Like human skin cells, the skin microbiome is continuously shed from its host and deposited on other individuals, items, and surfaces. For every one squamous epithelial cell shed from the human skin approximately 30 microorganisms are shed (5). Thus, deposited microorganisms could serve as an additional source of evidence to include or exclude a person of interest in criminal cases.

Genetic signatures from the skin microbiome could be used for HID with a panel that targets stable and abundant microorganisms (6, 7). Schmedes et al. (8) developed a targeted genome sequencing (TGS) panel called hidSkinPlex. This panel contains 286 markers covering a range of taxonomies of specific microorganisms that are in high abundance on the human skin (9). With the greater resolution of the hidSkinPlex and the demonstrated stability of the microorganisms chosen for the panel, a new avenue is available for HID using the skin microbiome. Although some relatively high accuracies were obtained the hidSkinPlex has areas that can be improved (10). Optimization of the number and informativeness of the markers as well as reduction in their amplicon size is still needed to improve the robustness for HID purposes.

Like the process of selecting ancestry informative markers in humans, single nucleotide polymorphisms (SNPs) within the hidSkinPlex panel of markers can be selected for HID using Wright's fixation index (F_{ST}). F_{ST} is an estimate of population differentiation that can be used to select SNPs to potentially increase classification accuracies for HID. Sherier et al. (11) used F_{ST} to select SNPs for HID; however the approach only focused on SNPs that were common to the two samples being compared. One ramification of only using SNPs that are local to a pair of

individuals is losing information that is specific in one individual but missing in the other. A global panel may allow for a better population genetic characterization of the selected SNPs. Furthermore, reducing to a specific set of informative SNPs can improve upon the efficiency of machine learning. Also, defined SNPs allow for better primer design for smaller amplicons which in turn can improve amplification efficiency (i.e., increased sensitivity of detection). Overall, HID could be easier to accomplish when the metrics for comparison are the same among all individuals.

The study herein focuses on developing a select microbial SNP panel for HID that is highly effective at associating a sample with its host. An effective microbial SNP panel would be well-defined in nearly all individuals, be highly individualizing, and involve typing as few genetic markers as possible to achieve a defined level of attribution. One approach to select specific SNPs and define potential accuracy is to consider the human host as a class and to leverage classifier algorithms to predict the identity of the human host from microbial signatures. Some classification algorithms can be used to learn a sparse solution (i.e., few SNP markers), and the data can be described in a way that is robust to missing data, a common problem with forensic samples. The least absolute shrinkage and selection operator (LASSO) is one such algorithm which can be used to simultaneously identify a sparse set of SNPs and use those SNPs for HID. Herein, a machine learning procedure is introduced that tests the ability of LASSO to identify SNPs for microbial-based HID. The performance of the selected SNP panel is then tested on individuals not used to generate the panel using a cross-validation framework. The candidate SNPs were assessed by their ability to predict the human host.

RESULTS

F_{ST} Estimation. The microbiomes of 51 individuals sampled at three body-sites in triplicate (total samples, n = 459) were sequenced with the hidSkinPlex panel. As described in Woerner et al. (10),

the resulting fastq files were aligned to the metagenomics database of MetaPhlAn2 (12). Wright's fixation index (F_{ST}) was estimated between all pairs of samples ($\binom{459}{2} = 105{,}111$ pairs) using the formulation of Hudson et al. (13) as described in Sherier et al. (11).

There are two objectives for determining single nucleotide markers for HID. The SNPs should be 1) individualizing to the person (i.e., have generally high F_{ST}) and 2) relatively stable over time. In terms of the first objective, F_{ST} varies from zero to one with an estimate of one indicating complete differentiation between (microbial) populations; thus, the nucleotides selected should tend to have large F_{ST}. In terms of the second objective, F_{ST} is undefined in the presence of missing data and when the allele is monomorphic between (and within) populations. Thus, it follows that the sites selected should have a F_{ST} that tends to be well defined. To evaluate the interplay between missing information and the central tendencies of the F_{ST} of microbial markers, F_{ST} was estimated at all the 172,116 nucleotide positions in the hidSkinPlex in anywhere from 1 to $\binom{459}{2} = 105{,}111$ pairwise comparisons (as limited by the information apparent, Figure 1). Nucleotide positions with F_{ST} estimates ≥ 0.1 and defined in at least 75% of the comparisons are herein considered candidate SNPs. Additionally, selecting SNPs seen in at least 75% of the pairwise comparisons allow for some tolerance for missing data which may be due to technical limitations.

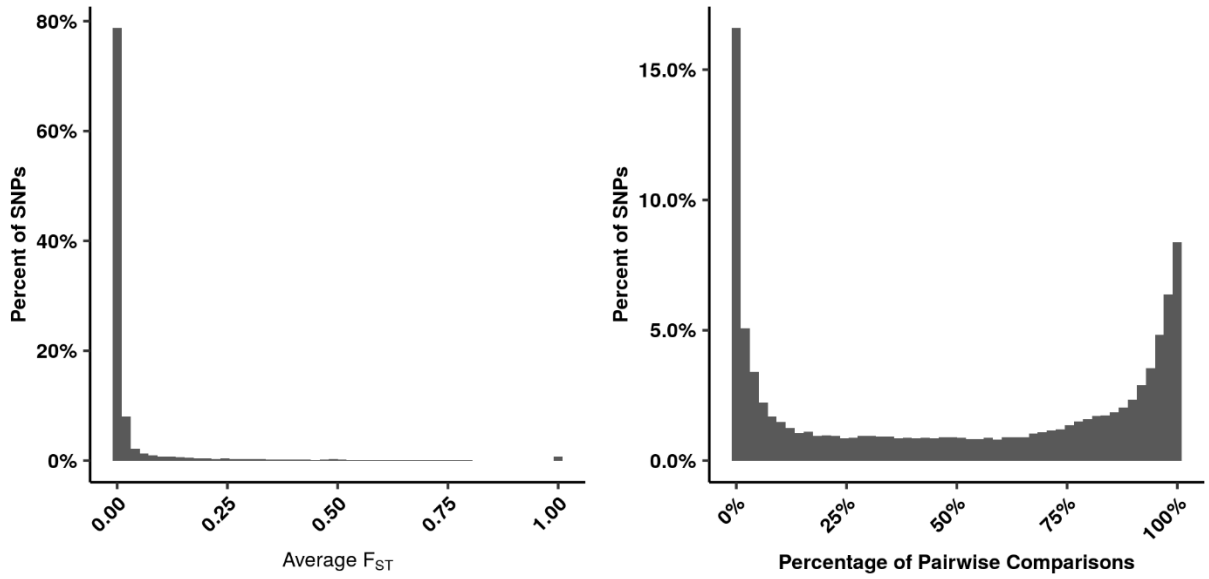


Figure 1. The average F_{ST} estimate and the sample size in the hidSkinPlex. The figure on the left shows the distribution of the average F_{ST} for all nucleotide positions in the hidSkinPlex. The figure on the right shows the percentage of nucleotide positions in which F_{ST} can be estimated.

Training and test data set creation. Training (n = 234, 26 individuals at three body-sites in triplicate) and test (n = 225, 25 individuals at three body-sites in triplicate) data were randomly partitioned (Supplemental S1). The training data set produced 26 SNP panels (one for each individual) with a mean of 1,265.769 ± 21.486 SNPs per panel. Similarly, the test data set produced 25 panels with a mean of 1,475.240 ± 45.256 SNPs per panel. For the final panel using all 459 samples from the training and test data set, the list of initial SNPs for analysis contained 4,445 SNPs (Figure 2).

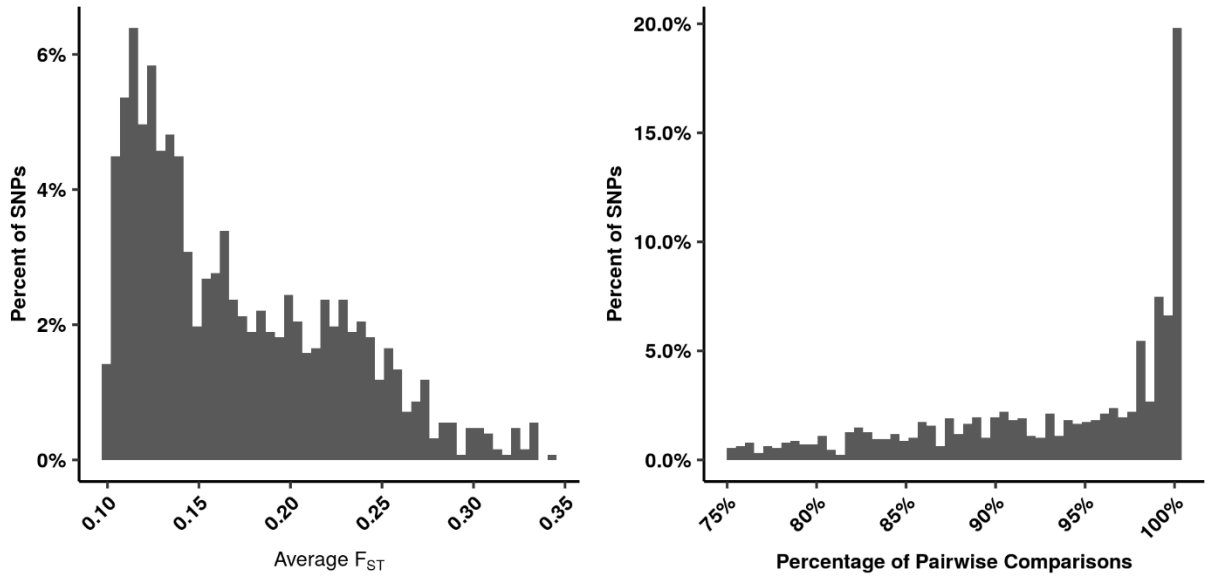


Figure 2. The average F_{ST} estimate and the sample size of the reduced list of 1,344 candidate SNPs from the training data set. The figure of the left shows the distribution of the average F_{ST} estimated for the SNP candidate list. The figure on the right shows the distribution of SNPs contained in the top 75% of pairwise comparisons.

Analysis of the training dataset. The training data set was used to optimize an algorithm for selecting a reduced number of SNPs for HID. The lambda sequence and the alpha parameter were optimized (see **Materials and Methods**), using all 26 individuals, to ensure that there were not too few or too many SNPs selected. A procedure was developed to select a SNP panel and then classify a new individual based on the selected markers. The procedure was run in a cross-validation framework, holding out each individual in turn. The training data set produced 26 SNP panels with a mean of 191.400 ± 21.702 SNPs.

Classification results. The above approach for selecting a reduced SNP list was applied to the training and test data sets. Applying the classification procedure (see **Materials and Methods**) to the training data set gave an overall accuracy of 93% (MCC = 0.920, 24.180 times better than chance), with only 18 out of 234 samples incorrectly classified (Figure 3). Of the incorrectly classified samples, four samples were from the foot (Fb), 11 from the manubrium (Mb), and three

from the hand (Hp) (Table 1). Samples from the Mb had a higher number of incorrectly classified samples compared to the Fb and a significantly higher number than the Hp (Fisher's Exact Test, p = 0.101 and 0.0470, respectively). A missing SNP was defined as a site selected in the procedure wherein 0 reads were apparent for a given sample. For the training data, the number of samples missing SNPs was determined for each of the 26 SNP panels. The mean number of samples missing SNPs was 93.300 ± 7.394 per panel. The mean number of missing SNPs for combined predicted results was 5.154 ± 11.312. For correctly classified samples, 131 samples were not missing any SNPs and 85 had missing SNPs with a mean of 10.330 ± 13.342. Fourteen incorrectly classified samples were missing SNPs with a mean of 23.430 ± 18.241. Incorrectly classified samples were more likely to have missing SNPs than correctly classified samples (Fisher's Exact Test, p = 0.002). The held-out sample for the development of each panel is more likely to have missing data because the SNPs were selected without considering the held out sample.

Table 1. The classification accuracy at different body sites in the training data set.

Training	*Foot (Fb)*	*Manubrium (Mb)*	*Hand (Hp)*	*Total*
Correct	74 (95%)	67 (85%)	75 (96%)	216 (93%)
Incorrect	4 (5%)	11 (15%)	3 (4%)	18 (7%)
Total	78	78	78	234

The classification procedure was applied to the test data set. The test data set was 96% accurate (MCC 0.954, 24.000 times better than chance), with only 10 (all from the Mb) out of 225 samples incorrectly classified (Table 2). Nine misclassified samples were missing a mean of 23.333 ± 32.943 SNPs (Figure 3). Of the correctly classified samples 103 out of 215 had missing SNPs (6.786 ± 8.354). As with the training data set, incorrectly classified samples were more likely to have missing SNPs than correctly classified samples (Fisher's Exact Test, p = 0.009).

Table 2. The classification accuracy at different body sites in the test data set.

Test	*Foot (Fb)*	*Manubrium (Mb)*	*Hand (Hp)*	*Total*
Correct	75 (100%)	65 (87%)	75 (100%)	215 (96%)
Incorrect	0 (0%)	10 (13%)	0 (0%)	10 (4%)
Total	75	75	75	225

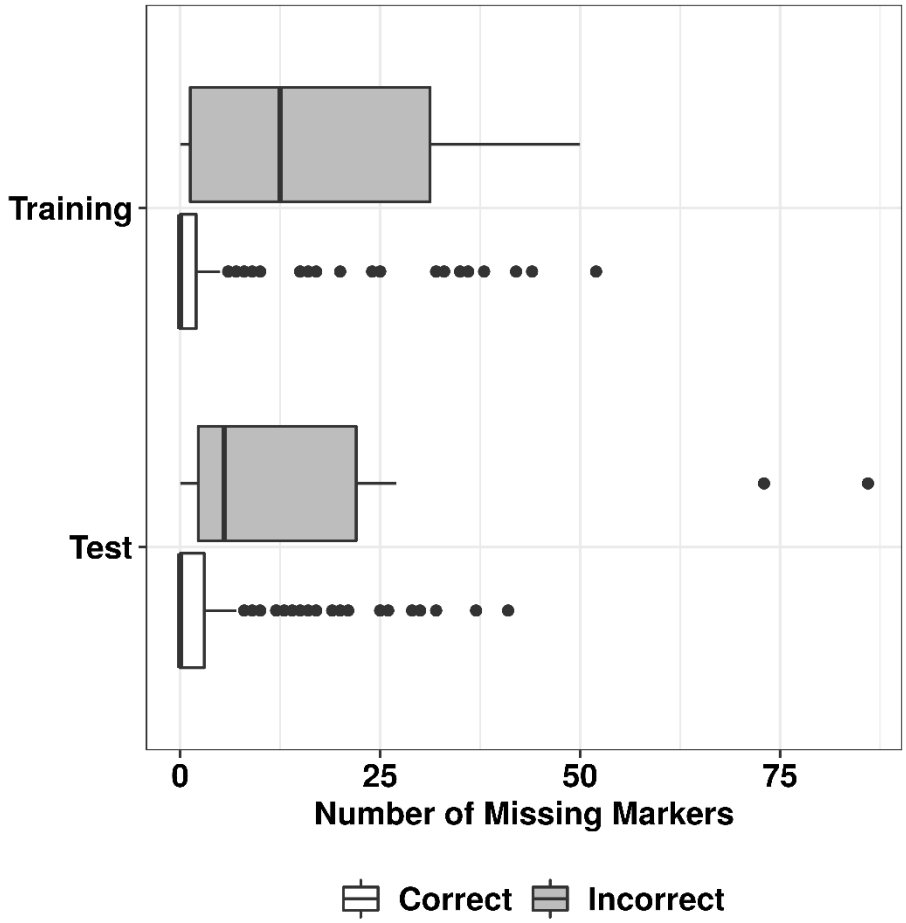


Figure 3. Classification results for training and test data sets and the number of samples missing SNPs. The x-axis indicates the number of missing SNPs for a given sample. The y-axis shows training and test data sets partitioned into the correct (white) and incorrect (gray) classification groups.

Reduced SNP list. The final candidate SNP list was determined by pooling the test and training data sets and re-applying a similar classification procedure. LASSO was used to produce a single SNP list. Cross validation was used to find the optimal lambda and to estimate the overall accuracy. The final SNP list, referred to as hidSkinPlex+, is composed of 365 SNPs (Supplemental Table

S2) that reside in 135 of the original amplicons (mean number of SNPs in each marker 3.419 ± 4.984, range from 1 to 51) from the hidSkinPlex (12). The markers are specific to four taxa, *Cutibacterium acnes*, *Cutibacterium humerusii, Corynebacterium tuberculostearicum, and Propionibacteriaceae*. Previous studies have shown *Cutibacterium* is a common and abundant (14-16) genus found on human skin (3, 9).

Of 459 samples, 95% (MCC = 0.949, 48.469 times better than chance) were correctly classified using data from the hidSkinPlex+. Of the 23 incorrectly classified samples, 17 were from Fb samples, which is a significantly larger number of samples than the two Mb and the four Hp samples incorrectly classified (Fisher's Exact Test, $p < 0.001$ compared to the Mb, $p = 0.003$ compared to Hp). The number of Mb and Hp was not significantly different ($p = 0.684$; Table 3). Of the 23 incorrectly classified samples 21 samples were missing a mean of 40.520 ± 36.853 SNPs. More incorrectly classified Fb samples had missing SNPs ($p < 0.001$) compared to the number of incorrectly classified Mb or Hp samples with missing SNPs. For the 436 samples correctly classified, 204 samples had a mean of 11.590 ± 16.559 of missing SNPs. A sample was more likely to be misclassified if it had missing SNPs (Fisher's Exact Test, $p < 0.001$).

Table 3. The classification accuracy at different body sites for hidSkinPlex+.

All Data	*Foot (Fb)*	*Manubrium (Mb)*	*Hand (Hp)*	*Total*
Correct	136 (89%)	151 (99%)	149 (97%)	436 (95%)
Incorrect	17 (11%)	2 (1%)	4 (3%)	23 (5%)
Total	153	153	153	459

DISCUSSION

The skin microbiome is a highly abundant and relatively stable source of DNA that may be utilized for HID (6, 17-21). A common set of microbial SNPs could provide another avenue of investigation to improve HID. In this study, a subset of SNPs from the hidSkinPlex that generally

were common to all individuals analyzed were assessed for classification accuracy. The skin microbiome samples from 51 individuals' Fb, Mb and Hp were attributed to their respective individual hosts with an accuracy of 96% for the test data set. The targeted panels were composed of 157 to 243 SNPs, a substantial decrease in the number of SNPs relied on by Woerner et al. (10). The final SNP panel, the hidSkinPlex+, contained 365 SNPs residing in 135 markers which were specific to four taxa. LASSO was used to select informative SNPs for HID and correctly predicted the human host 95% of the time. It should be noted, however, that reported accuracy of the final panel may be slightly biased upwards as it is estimated within-fold, though given the 96% accuracy of the test data set this bias is likely modest. Classification accuracies for each of the three body sites using the hidSkinPlex+ ranged from 89 – 99%. Accuracy with Fb samples (89%) was significantly lower (i.e., a greater number of incorrectly classified samples) compared with the Mb (99%, Fisher's Exact Test $p < 0.001$) and Hp (97%, Fisher's Exact Test $p = 0.001$) samples, while the accuracy for the Mb and Hp sites were not significantly different. While accuracies for the Fb in this study were lower than the other two body sites, the results were more accurate than previous work from Woerner et al. (10) (28% - 73%).

One factor that appears to be related to the reduced accuracies is missing SNPs. Samples that had missing SNPs were more likely to be incorrectly classified compared to samples that had no missing SNPs (Fisher's Exact Test, $p < 0.001$ for all comparisons). However, there were also samples that had missing SNPs that were classified correctly. For example, 38 out of 149 Hp samples had missing SNPs but were correctly classified. The incorrectly classified Hp samples had a significantly higher mean number of missing SNPs (34.500 ± 32.296) compared to Hp samples with missing data that were correctly classified (13.340 ± 16.515, Fisher's Exact Test $p < 0.001$).While further research is needed to determine why some samples were incorrectly

classified, one possible explanation of incorrect classification is low coverage. The Fb had the largest amount of missing data and the lowest read coverage. For example, the Fb had a mean of 448,400 ± 319,227 reads compared to Hp which had a mean of 1,025,866 ± 410,674 reads. Therefore, a more efficient chemistry could reduce the chances of data drop out.

The hidSkinPlex+ allows for a new targeted sequencing panel to be designed and optimized for the use of HID. Eliminating the markers that do not contribute to classification accuracy can improve the enrichment process, i.e., amplification efficiency of the polymerase chain reaction (PCR). Fewer markers in a PCR may increase amplicon yield and thus provide a more sensitive assay. Since the hidSkinPlex+ contains fewer markers, and thus SNPs, than that of the original hidSkinPlex, specific targeted SNPs primers may be redesigned to generate smaller amplicons that may increase amplicon yield and provide for a more robust panel for analyzing degraded samples, which are desirable features for forensic applications. Research is still needed to assess how well the SNPs selected for the hidSkinPlex+ work when applied to samples collected from touch samples and at different time points. With additional studies on the allele frequency of the selected SNPs in different populations (populations may be geographically determined instead of genetically determined) a better estimation of HID classification accuracies can be achieved.

These results, herein, further support the skin microbiome can serve as a potential source of DNA for HID. This panel could serve as a set of biomarkers to assess the stability of the specific SNPs and whether they can be generalized to the greater population.

MATERIALS AND METHODS

Sample collection and sequencing. Human skin microbiome samples were collected by swabbing 51 individuals at three body sites (manubrium (Mb), hand (Hp), foot (Fb)) in triplicate (replicates R1, R2, R3), for a total of 459 samples as described previously in Woerner et al. (10). Briefly, the

samples were assayed with the hidSkinPlex, a TGS panel developed by Schmedes et al. (9). All markers in the hidSkinPlex (9) are drawn from the MetaPhlAn2 database (12) and as such describe both the nucleotide sequence of the marker as well as a corresponding taxonomic affiliation (e.g., the marker is associated with *C. acnes*). The hidSkinPlex panel targets 22 clades from the genus to the species level and is comprised of 286 markers that are considered taxonomically stable and abundant on human skin (9). The University of North Texas Health Science Center Institutional Review Board approved the collection and analyses of these samples.

Sequence data generated. As described previously in Woerner et al. (10), all sequencing was performed on a MiSeq (Illumina, San Diego, CA). Fastq files were trimmed with cutadapt (22) to remove adapters from the sequencing results. The sequence data were aligned to the MetaPhlAn2 reference database (12) using bowtie2. Using an in-house BASH script (v. 4.4.20, Free Software Foundation, http://www.gnu.org/software/bash/), the total number of reads and the percent of ACGT for each nucleotide base were calculated based on pileups from samtools (23). Finally, the base pileups for each aligned marker in the hidSkinPlex panel were generated.

Computation and statistical analysis. As described in Sherier et al. (11), the F_{ST} was computed as per Hudson et al. (13), who proposed estimating F_{ST} as $F_{ST} = 1 - (H_w/H_b)$, where H_w is the mean number of pairwise differences within a population and H_b is the mean number of pairwise differences between two populations (13). F_{ST} was estimated using an in-house script written in the Python programming language (v. 2.7.17, Python Software Foundation, https:/www.python.org/) with minor modifications from the script used in Sherier et al. (11). The modifications allowed F_{ST} to be measured at all nucleotide positions regardless of read depth. All other statistical analyses were performed in R (v. 4.0.3) using the glmnet package (v4.1-1) (24),

the tidyverse (v. 1.3.1) (25), and ggplot2 (v. 1.3.1) (26) as appropriate. Additionally, Matthews correlation coefficient (MCC) was estimated using mltools (v. 0.3.5) (27).

Potential SNPs for analysis. Potential informative nucleotide positions were identified on the basis of F_{ST} estimated between pairs of samples. Training (n = 234, 26 individuals at three body-sites in triplicate) and test (n = 225, 25 individuals at three body-sites in triplicate) data were randomly partitioned by `sample(c(1:51), 26)` in R (Supplemental S1). F_{ST} was estimated between all samples within the training and the test data sets separately (27,261 and 25,200 pairwise comparisons respectively) at all 172,116 nucleotide positions in the hidSkinPlex. As a summary statistic, F_{ST} is undefined if either individual (or both individuals) is missing the SNP or if the allele is monomorphic between populations. Thus, for some pairwise comparisons there is no F_{ST} estimate. F_{ST} estimates less than zero were documented as zero. Herein, candidate SNPs were defined as nucleotide positions that have a mean F_{ST} estimate ≥ 0.1 and had a defined F_{ST} in > 75% of comparisons.

Machine learning strategy. A major aim of the current study is to identify SNPs that are informative for HID. The SNPs are selected to both differentiate individuals (e.g., tend to have high F_{ST}) and to be well-defined (have a defined F_{ST}). Further, a central aim is to identify a small number of such SNPs (i.e., a sparse solution). Classification algorithms can be used for HID by treating each person as a class (i.e., as a categorical variable) and for the approaches herein an individual is predicted based on coefficients that are learned for that individual. One tool to find a small set of SNPs for HID is LASSO (the least absolute shrinkage and selection estimator). LASSO considers two measures in its optimization: the error (i.e., deviance in a logistic regression) and the absolute value of the coefficients (i.e., the L_1 norm). The relative importance of these two criteria is specified with lambda. L_1 regularization tends to produce solutions that are sparse. Thus,

in the current use-case LASSO can be used to simultaneously identify a SNP panel and predict the human host based on those SNPs.

A potential concern with LASSO is that the SNP panel identified may work well for each person currently in the database, however it may not work well for new individuals. If one were to consider a potential forensic scenario, a new individual (e.g., a person of interest) is presented and the SNP panel needs to both be accurate for the current database and the additional individual (i.e., the new class). While the collection of the samples used in this study does not mimic real-life casework, the classification method should determine if accurate HID is possible when samples are technical replicates and collected directly from an individual. Given these requirements, a procedure was developed that first learns a sparse set of SNPs using LASSO from individuals in a database, a new individual is introduced, and then the additional individual is classified based on a SNP panel. The last classification is performed using a ridge regression (L_2 norm) with the SNP panel developed within fold for each held out individual (i.e., based on high F_{ST} SNPs identified in the database and not considering the held out individual).The three-step procedure was repeated in a cross-validation framework, holding out each individual in turn. To ensure that the sample sizes were equal in all classes during the cross-validation development of the reduced SNP panel, the same sample-type (e.g., Hp, replicate 3) was held out in all individuals. In R (28), the lambda sequence is given by 1.1^seq(1, -200, length=100). SNP coefficients and the optimal lambda value were learned using the `cv.glmnet` function in the R package `glmnet`. The optimal lambda was taken to be the lambda that minimizes the deviance (`lambda.min` from `cv.glmnet`). In particular, the LASSO regression was run by standardizing the allele frequencies for the provided SNPs (`standardize = TRUE`), the regression type was set to grouped (`type.multinomial = "grouped"`), and the maximum number of iterations were set to

1,000,000. SNPs were identified by selecting SNP alleles based on the optimal lambda from LASSO. SNP panels were created by using all allele frequencies for any SNP position corresponding to a non-zero coefficient. A ridge regression was used to predict the held-out individual (that is, by setting `alpha=0` in `cv.glmnet`, but otherwise as per the above LASSO procedure).

Selection of SNPs for hidSkinPlex+. The machine learning strategy above was designed to simulate the ability of LASSO to identify SNPs in some data set that can then be used to predict the identity of previously unseen individual. In the framework above a SNP panel is produced for each held-out individual, which is appropriate for assessing the accuracy of the approach, but it does not create a singular SNP panel. To produce a final SNP panel a similar but simpler procedure was used. LASSO was used to identify a sparse set of SNPs considering all individuals (and body sites) pooled across the training and test data sets. The same lambda sequence was considered and the optimal lambda (and corresponding panel) was estimated using cross-validation (`cv.glmnet`). The final panel is referred to as the hidSkinPlex+. The accuracy of the final panel was estimated within-fold (`keep = TRUE`), and, as such, the estimated accuracy of the final panel is likely inflated (biased upwards).

DATA AVAILABILITY

Custom R and Python scripts can be accessed at https://github.com/CardiShire/MLforSkinMicrobiomeHID.

ACKNOWLEDGEMENTS

We thank Sarah Schmedes for the design of the hidSkinPlex and sample processing. Additionally, we thank Angie Ambers, Rachel Kieser, Frank Wendt, Nicole Novroski, and Jonathan King for their contributions to collecting/processing samples. We also would like to thank Utpal Smart, Sammed Mandape, Ben Crysup, and Jonathan King for all the time they spent advising on code and debugging.

This study was supported in part by the National Institute of Justice, award numbers 2015-NE-BX-K006 and 2020-R2-CX-0046. The views expressed in this article do not necessarily represent the views of the Department of Justice, National Institute of Justice, or the United States government.

BIBLIOGRAPHY

1. Wang Y, Yu Q, Zhou R, Feng T, Hilal MG, Li H. 2021. Nationality and body location alter human skin microbiome. Applied Microbiology and Biotechnology doi:10.1007/s00253-021-11387-8.
2. Ross AA, Doxey AC, Neufeld JD. 2017. The Skin Microbiome of Cohabiting Couples. mSystems 2:e00043-17.
3. Oh J, Byrd AL, Deming C, Conlan S, Program NCS, Kong HH, Segre JA. 2014. Biogeography and individuality shape function in the human skin metagenome. Nature 514:59-64.
4. Richardson M, Gottel N, Gilbert JA, Lax S. 2019. Microbial Similarity between Students in a Common Dormitory Environment Reveals the Forensic Potential of Individual Microbial Signatures. mBio 10:e01054-19.
5. Percival SL, Emanuel C, Cutting KF, Williams DW. 2012. Microbiology of the skin and the role of biofilms in infection. International Wound Journal 9:14-32.
6. 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 107:6477-81.
7. Knight R, Metcalf JL, Gilbert JA, Carter DO. 2018. Evaluating the Skin Microbiome as Trace Evidence. National Criminal Justice Reference Service.
8. Oh J, Byrd AL, Park M, Program NCS, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854-66.
9. Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. 2018. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Sci Int Genet 32:50-61.

10. Woerner AE, Novroski NMM, Wendt FR, Ambers A, Wiley R, Schmedes SE, Budowle B. 2019. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci Int Genet 38:130-139.

11. Sherier AJ, Woerner AE, Budowle B. 2021. Population Informative Markers Selected Using Wright's Fixation Index and Machine Learning Improves Human Identification Using the Skin Microbiome. Appl Environ Microbiol 87:e0120821.

12. 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 12:902-3.

13. Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-9.

14. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Program NCS, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190-2.

15. Grice EA, Kong HH, Renaud G, Young AC, Program NCS, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre JA. 2008. A diversity profile of the human skin microbiota. Genome Res 18:1043-50.

16. Fitz-Gibbon S, Tomida S, Chiu BH, Nguyen L, Du C, Liu M, Elashoff D, Erfe MC, Loncaric A, Kim J, Modlin RL, Miller JF, Sodergren E, Craft N, Weinstock GM, Li H. 2013. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. J Invest Dermatol 133:2152-60.

17. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930-8.

18. Hampton-Marcell JT, Larsen P, Anton T, Cralle L, Sangwan N, Lax S, Gottel N, Salas-Garcia M, Young C, Duncan G, Lopez JV, Gilbert JA. 2020. Detecting personal microbiota signatures at artificial crime scenes. Forensic Sci Int 313:110351.

19. Kapono CA, Morton JT, Bouslimani A, Melnik AV, Orlinsky K, Knaan TL, Garg N, Vazquez-Baeza Y, Protsyuk I, Janssen S, Zhu Q, Alexandrov T, Smarr L, Knight R, Dorrestein PC. 2018. Creating a 3D microbial and chemical snapshot of a human habitat. Sci Rep 8:3669.

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

21. Lee S-Y, Woo S-K, Lee S-M, Eom Y-B. 2016. Forensic analysis using microbial community between skin bacteria and fabrics. Toxicology and Environmental Health Sciences 8:263-270.

22. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:3.

23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.

24. Friedman J, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33:1--22.

25. Wickham H, Averick M, Bryan J, Chang W, McGowan LDAF, Romain , Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the {tidyverse}. Journal of Open Source Software 4:1686.

26. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K. 2016. ggplot2: Elegant Graphics for Data Analysis, vol 2018. Springer-Verlag New York.

27. Gorman B. 2018. Package 'mltools'. https://github.com/ben519/mltools. Accessed

28. Team RC. 2013. R: A Language and Environment for Statistical Computing, *on* R Foundation for Statistical Computing. http://www.R-project.org/. Accessed 2/12/21.

SUPPLEMENTAL

S1 List of samples contained in the training and test data set.

Training

1, 2, 4, 6, 7, 10, 11, 12, 15, 16, 17, 25, 28, 30, 31, 32, 33, 38, 40, 42, 44, 45, 46, 47, 49, 51

Test

3, 5, 8, 9, 13, 14, 18, 19, 20, 21, 22, 23, 24, 26, 27, 29, 34, 35, 36, 37, 39, 41, 43, 48, 50

S2 List of SNPs selected for hidSkinPlex+.

Marker	SNP
NC_014039.1:c1439020-1438442	111
NC_014039.1:c1439020-1438442	120
NC_014039.1:c1439020-1438442	149
NC_014039.1:c1439020-1438442	162
NC_014039.1:c1439020-1438442	169
NC_014039.1:c1439020-1438442	339
NC_014039.1:c1439020-1438442	369
NC_014039.1:c1439020-1438442	43
NC_016511.1:2485446-2486162	325
NC_016511.1:2485446-2486162	47
NC_016511.1:2485446-2486162	613
NC_017535.1:c1339878-1339075	597
NC_018707.1:c1315368-1314979	313
NZ_ACVP01000023.1:c144466-143744	147
NZ_ACVP01000023.1:c144466-143744	156
NZ_ACVP01000023.1:c144466-143744	171
NZ_ACVP01000023.1:c144466-143744	177
NZ_ACVP01000023.1:c144466-143744	178
NZ_ACVP01000023.1:c144466-143744	183
NZ_ACVP01000023.1:c144466-143744	216
NZ_ACVP01000023.1:c144466-143744	225
NZ_ACVP01000023.1:c144466-143744	226
NZ_ACVP01000023.1:c144466-143744	231
NZ_ACVP01000023.1:c144466-143744	249
NZ_ACVP01000023.1:c144466-143744	255
NZ_ACVP01000023.1:c144466-143744	262
NZ_ACVP01000023.1:c144466-143744	273

NZ_ACVP01000023.1:c144466-143744	276
NZ_ACVP01000023.1:c144466-143744	279
NZ_ACVP01000023.1:c144466-143744	285
NZ_ACVP01000023.1:c144466-143744	303
NZ_ACVP01000023.1:c144466-143744	312
NZ_ACVP01000023.1:c144466-143744	351
NZ_ACVP01000023.1:c144466-143744	366
NZ_ACVP01000023.1:c144466-143744	384
NZ_ACVP01000023.1:c144466-143744	393
NZ_ACVP01000023.1:c144466-143744	402
NZ_ACVP01000023.1:c144466-143744	414
NZ_ACVP01000023.1:c144466-143744	417
NZ_ACVP01000023.1:c144466-143744	423
NZ_ACVP01000023.1:c144466-143744	432
NZ_ACVP01000023.1:c144466-143744	448
NZ_ACVP01000023.1:c144466-143744	451
NZ_ACVP01000023.1:c144466-143744	486
NZ_ACVP01000023.1:c144466-143744	520
NZ_ACVP01000023.1:c144466-143744	555
NZ_ACVP01000023.1:c144466-143744	577
NZ_ACVP01000023.1:c144466-143744	613
NZ_AFAM01000001.1:c260639-259980	363
NZ_AFAM01000001.1:c260639-259980	380
NZ_AFAM01000001.1:c260639-259980	381
NZ_AFAM01000001.1:c260639-259980	396
NZ_AFAM01000001.1:c260639-259980	423
NZ_AFAM01000001.1:c260639-259980	426
NZ_AFAM01000001.1:c260639-259980	455
NZ_AFAM01000001.1:c260639-259980	489
NZ_AFAM01000001.1:c260639-259980	519
NZ_AFAM01000005.1:c52756-52631	103
NZ_AFAM01000005.1:c52756-52631	113
NZ_AFAM01000005.1:c52756-52631	125
NZ_AFAM01000005.1:c52756-52631	71
NZ_AFAM01000005.1:c52756-52631	99
NZ_AFAM01000014.1:c59116-58358	422
NZ_AFAM01000014.1:c59116-58358	423
NZ_AFAM01000020.1:c4555-4424	42
NZ_AFAM01000020.1:c4555-4424	74
NZ_AFAM01000020.1:c4555-4424	75
NZ_AFIK01000013.1:c12739-12119	532
NZ_AFIK01000014.1:315-1133	318
NZ_AFIK01000014.1:315-1133	517

NZ_AFIK01000020.1:c12439-12299	91
NZ_AFIK01000020.1:c12439-12299	92
NZ_AFIK01000053.1:c36245-34977	462
NZ_AFIK01000053.1:c36245-34977	657
NZ_AFIK01000065.1:c4330-4001	192
NZ_AFIK01000082.1:c111360-110575	610
NZ_AFIK01000082.1:c111360-110575	700
NZ_AFIL01000010.1:c43071-42837	209
NZ_AFIL01000010.1:c43071-42837	37
NZ_AFIL01000010.1:c43071-42837	56
NZ_AFIL01000010.1:c43071-42837	79
NZ_AFIL01000016.1:c75436-75296	141
NZ_AFIL01000025.1:23315-23623	192
NZ_AFIL01000030.1:c58004-57372	114
NZ_AFIL01000031.1:46041-46637	268
NZ_AFIL01000040.1:4048-4263	133
NZ_AFIL01000040.1:4048-4263	89
NZ_AFIL01000041.1:c77880-77749	104
NZ_AFIL01000041.1:c77880-77749	112
NZ_AFIL01000041.1:c77880-77749	57
NZ_AFIL01000041.1:c77880-77749	74
NZ_AFIL01000047.1:12103-12642	195
NZ_AFIL01000051.1:c25042-24929	71
NZ_AFIL01000069.1:c9632-8838	25
NZ_AFIL01000069.1:c9632-8838	308
NZ_AFIL01000070.1:3643-4386	620
NZ_AFUK01000001.1:1588290-1589009	573
NZ_AFUK01000001.1:1828645-1829349	502
NZ_AFUK01000001.1:527724-528653	324
NZ_AFUK01000001.1:527724-528653	369
NZ_AFUK01000001.1:527724-528653	621
NZ_AFUK01000001.1:527724-528653	678
NZ_AFUK01000001.1:535213-535428	135
NZ_AFUK01000001.1:535213-535428	33
NZ_AFUK01000001.1:535213-535428	82
NZ_AFUK01000001.1:665124-666446	1071
NZ_AFUK01000001.1:665124-666446	1181
NZ_AFUK01000001.1:665124-666446	549
NZ_AFUK01000001.1:665124-666446	600
NZ_AFUK01000001.1:c1255510-1255055	290
NZ_AFUK01000001.1:c1579497-1578787	292
NZ_AFUK01000001.1:c1579497-1578787	400
NZ_AFUK01000001.1:c1715790-1715233	202

NZ_AFUK01000001.1:c1845075-1844710	160
NZ_AFUK01000001.1:c359834-359544	147
NZ_AFUK01000001.1:c359834-359544	189
NZ_AFUK01000001.1:c359834-359544	216
NZ_AXME01000001.1:1088727-1089377	504
NZ_AXME01000001.1:1146402-1146932	155
NZ_AXME01000001.1:1146402-1146932	165
NZ_AXME01000001.1:1286960-1287442	66
NZ_AXME01000001.1:1327950-1328573	354
NZ_AXME01000001.1:1431752-1431913	61
NZ_AXME01000001.1:1431752-1431913	62
NZ_AXME01000001.1:1431752-1431913	92
NZ_AXME01000001.1:40840-41742	120
NZ_AXME01000001.1:40840-41742	161
NZ_AXME01000001.1:40840-41742	267
NZ_AXME01000001.1:40840-41742	324
NZ_AXME01000001.1:40840-41742	498
NZ_AXME01000001.1:40840-41742	647
NZ_AXME01000001.1:49241-49654	219
NZ_AXME01000001.1:49241-49654	285
NZ_AXME01000001.1:655649-655855	103
NZ_AXME01000001.1:655649-655855	142
NZ_AXME01000001.1:655649-655855	58
NZ_AXME01000001.1:702826-703131	130
NZ_AXME01000001.1:97330-98208	639
NZ_AXME01000001.1:97330-98208	693
NZ_AXME01000001.1:97330-98208	720
NZ_AXME01000001.1:990664-990933	147
NZ_AXME01000001.1:990664-990933	149
NZ_AXME01000001.1:990664-990933	197
NZ_AXME01000001.1:990664-990933	89
NZ_AXME01000001.1:c1552174-1551533	256
NZ_AXME01000001.1:c1599141-1598893	93
NZ_AXME01000001.1:c1820429-1820292	53
NZ_AXME01000001.1:c1820429-1820292	67
NZ_AXME01000001.1:c1820429-1820292	69
NZ_AXME01000001.1:c1820429-1820292	73
NZ_AXME01000001.1:c2014536-2014075	411
NZ_AXME01000001.1:c2135959-2134715	789
NZ_AXME01000001.1:c2447430-2446870	136
NZ_AXME01000001.1:c2447430-2446870	359
NZ_AXME01000001.1:c31864-31571	165
NZ_AXME01000001.1:c31864-31571	170

NZ_AXME01000001.1:c31864-31571	98
NZ_AXMG01000001.1:1150303-1151070	308
NZ_AXMG01000001.1:1231251-1231871	336
NZ_AXMG01000001.1:1231251-1231871	588
NZ_AXMG01000001.1:1440218-1440469	42
NZ_AXMG01000001.1:1440218-1440469	56
NZ_AXMG01000001.1:1440218-1440469	79
NZ_AXMG01000001.1:1440218-1440469	80
NZ_AXMG01000001.1:536557-537231	111
NZ_AXMG01000001.1:793445-793843	261
NZ_AXMG01000001.1:99114-99290	107
NZ_AXMG01000001.1:c1328090-1327596	314
NZ_AXMG01000001.1:c1443707-1443105	327
NZ_AXMG01000001.1:c1443707-1443105	333
NZ_AXMG01000001.1:c1443707-1443105	579
NZ_AXMG01000001.1:c1945194-1944973	30
NZ_AXMG01000001.1:c2126720-2126193	185
NZ_AXMG01000001.1:c2126720-2126193	318
NZ_AXMG01000001.1:c2126720-2126193	388
NZ_AXMG01000001.1:c2126720-2126193	389
NZ_AXMG01000001.1:c2126720-2126193	390
NZ_AXMG01000001.1:c2126720-2126193	501
NZ_AXMG01000001.1:c2312839-2311925	221
NZ_AXMG01000001.1:c2312839-2311925	281
NZ_AXMG01000001.1:c2312839-2311925	662
NZ_AXMG01000001.1:c2382295-2381897	120
NZ_AXMG01000001.1:c2382295-2381897	135
NZ_AXMG01000001.1:c2382295-2381897	140
NZ_AXMG01000001.1:c2382295-2381897	234
NZ_AXMG01000001.1:c2382295-2381897	354
NZ_AXMG01000001.1:c2382295-2381897	54
NZ_AXMG01000001.1:c2382295-2381897	84
NZ_AXMG01000001.1:c2429318-2428110	657
NZ_AXMG01000001.1:c2429318-2428110	696
NZ_AXMG01000001.1:c2429318-2428110	702
NZ_AXMI01000001.1:619555-620031	242
NZ_AXMI01000001.1:c101377-100163	400
NZ_AXMI01000001.1:c101377-100163	406
NZ_AXMI01000001.1:c101377-100163	407
NZ_AXMI01000001.1:c101377-100163	487
NZ_AXMI01000001.1:c101377-100163	816
NZ_AXMI01000001.1:c282323-281691	177
NZ_AXMI01000001.1:c282323-281691	180

NZ_AXMI01000001.1:c282323-281691	193
NZ_AXMI01000001.1:c282323-281691	214
NZ_AXMI01000001.1:c282323-281691	236
NZ_AXMI01000001.1:c282323-281691	260
NZ_AXMI01000001.1:c282323-281691	338
NZ_AXMI01000001.1:c282323-281691	343
NZ_AXMI01000001.1:c282323-281691	353
NZ_AXMI01000001.1:c282323-281691	441
NZ_AXMI01000001.1:c282323-281691	531
NZ_AXMI01000001.1:c282323-281691	558
NZ_AXMI01000001.1:c282323-281691	92
NZ_AXMI01000001.1:c282323-281691	97
NZ_AXMI01000001.1:c306684-306040	255
NZ_AXMI01000001.1:c306684-306040	318
NZ_AXMI01000001.1:c325088-324501	236
NZ_AXMI01000001.1:c443438-442323	135
NZ_AXMI01000001.1:c443438-442323	271
NZ_AXMI01000002.1:319095-319601	382
NZ_AXMI01000002.1:525312-525770	175
NZ_AXMI01000002.1:525312-525770	356
NZ_AXMI01000002.1:525312-525770	363
NZ_AXMI01000002.1:674988-675587	381
NZ_AXMI01000002.1:721564-722400	139
NZ_AXMI01000002.1:721564-722400	201
NZ_AXMI01000002.1:721564-722400	473
NZ_AXMI01000002.1:721564-722400	543
NZ_AXMI01000002.1:721564-722400	574
NZ_AXMI01000002.1:721564-722400	589
NZ_AXMI01000002.1:837080-837400	114
NZ_AXMI01000002.1:837080-837400	49
NZ_AXMI01000002.1:c247178-246402	420
NZ_AXMI01000002.1:c247178-246402	555
NZ_AXMI01000002.1:c247178-246402	631
NZ_AXMI01000002.1:c671938-670697	220
NZ_AXMI01000002.1:c671938-670697	656
NZ_AXMI01000002.1:c872629-871631	586
NZ_AXMI01000002.1:c872629-871631	655
NZ_AXMI01000003.1:232201-232740	172
NZ_AXMI01000003.1:232201-232740	76
NZ_AXMI01000004.1:13568-14401	105
NZ_AXMI01000004.1:13568-14401	627
NZ_AXMI01000004.1:13568-14401	648
NZ_AXMI01000004.1:48085-48816	239

NZ_AXMI01000004.1:48085-48816	426
NZ_AXMI01000004.1:c102788-101976	481
NZ_AXMI01000004.1:c102788-101976	588
NZ_AXMI01000004.1:c102788-101976	670
NZ_AXMI01000004.1:c231437-230883	219
NZ_AXMI01000004.1:c231437-230883	54
NZ_AXMI01000004.1:c577292-575922	1207
NZ_AXMI01000004.1:c577292-575922	1233
NZ_AXMI01000004.1:c577292-575922	618
NZ_AXMI01000004.1:c577292-575922	681
NZ_AXMI01000004.1:c577292-575922	719
NZ_AXMI01000004.1:c577292-575922	879
NZ_AXMI01000004.1:c577292-575922	987
NZ_AXMI01000006.1:1-107	77
NZ_AXMK01000001.1:c1228696-1228250	107
NZ_AXMK01000001.1:c1228696-1228250	284
NZ_AXMK01000001.1:c1228696-1228250	286
NZ_AXMK01000001.1:c1228696-1228250	338
NZ_AXML01000004.1:c579659-578172	115
NZ_AXML01000004.1:c579659-578172	153
NZ_AXML01000004.1:c579659-578172	162
NZ_AXML01000004.1:c579659-578172	249
NZ_AXML01000004.1:c579659-578172	273
NZ_AXML01000004.1:c579659-578172	315
NZ_AXML01000004.1:c579659-578172	345
NZ_AXML01000004.1:c579659-578172	402
NZ_AXML01000004.1:c579659-578172	430
NZ_AXML01000004.1:c579659-578172	435
NZ_AXML01000004.1:c579659-578172	477
NZ_AXML01000004.1:c579659-578172	542
NZ_AXML01000004.1:c579659-578172	96
NZ_GL383469.1:c216727-215501	1006
NZ_GL383469.1:c216727-215501	777
NZ_GL383469.1:c216727-215501	788
NZ_GL383714.1:170052-170369	297
NZ_GL383714.1:170052-170369	99
NZ_GL383759.1:c166532-166311	79
NZ_GL383759.1:c166532-166311	86
NZ_GL383759.1:c166532-166311	87
NZ_GL383802.1:56803-56916	35
NZ_GL383811.1:10443-11039	105
NZ_GL383811.1:10443-11039	255
NZ_GL383811.1:10443-11039	294

NZ_GL383811.1:10443-11039	408
NZ_GL383846.1:26161-26922	202
NZ_GL383846.1:26161-26922	222
NZ_GL383846.1:26161-26922	485
NZ_GL383846.1:26161-26922	522
NZ_GL383846.1:26161-26922	565
NZ_GL383846.1:26161-26922	566
NZ_GL383846.1:26161-26922	716
NZ_GL384462.1:c297812-297150	489
NZ_GL384610.1:c285619-284684	630
NZ_GL384610.1:c285619-284684	804
NZ_GL384611.1:c783227-783054	157
NZ_GL384611.1:c783227-783054	159
NZ_GL878448.1:c80834-80607	33
NZ_GL878448.1:c80834-80607	45
NZ_GL878455.1:c805995-805537	178
NZ_GL878455.1:c805995-805537	75
NZ_GL883048.1:64439-65218	578
NZ_JH376566.1:1103467-1104744	1044
NZ_JH376566.1:1103467-1104744	530
NZ_JH376566.1:1103467-1104744	678
NZ_JH376566.1:1103467-1104744	948
NZ_JH376566.1:1105369-1105965	58
NZ_JH376566.1:1105369-1105965	65
NZ_JH376566.1:326756-326986	210
NZ_JH376566.1:507019-507612	166
NZ_JH376566.1:882552-883256	296
NZ_JH376566.1:882552-883256	417
NZ_JH376566.1:882552-883256	571
NZ_JH376566.1:882552-883256	654
NZ_JH376567.1:190789-191232	407
NZ_JH376567.1:251291-251998	318
NZ_JH376567.1:251291-251998	417
NZ_JH376567.1:251291-251998	66
NZ_JH376567.1:251291-251998	84
NZ_JH376567.1:592116-592328	25
NZ_JH376567.1:598376-599065	555
NZ_JH376567.1:c388018-387605	138
NZ_JH376567.1:c388018-387605	69
NZ_JH376568.1:c255689-255105	461
NZ_JH376569.1:c80380-79448	537
NZ_JH376569.1:c80380-79448	575
NZ_JH376569.1:c80380-79448	783

NZ_KI515684.1:459339-460115	103
NZ_KI515684.1:459339-460115	315
NZ_KI515684.1:c325537-325361	150
NZ_KI515684.1:c44215-43715	276
NZ_KI515684.1:c44215-43715	61
NZ_KI515684.1:c488989-488798	105
NZ_KI515684.1:c488989-488798	93
NZ_KI515684.1:c584270-583890	116
NZ_KI515684.1:c96934-96368	306
NZ_KI515684.1:c96934-96368	517
NZ_KI515684.1:c96934-96368	518
NZ_KI515685.1:1081256-1081411	148
NZ_KI515685.1:187493-188140	411
NZ_KI515685.1:187493-188140	548
NZ_KI515685.1:225601-226386	636
NZ_KI515685.1:339623-340705	603
NZ_KI515685.1:339623-340705	82
NZ_KI515685.1:432422-433465	305
NZ_KI515685.1:546580-547218	453
NZ_KI515685.1:c1032381-1030873	347
NZ_KI515685.1:c157510-157292	18
NZ_KI515685.1:c743399-743001	129
NZ_KI515685.1:c743399-743001	235
NZ_KI515685.1:c849089-848304	229
NZ_KI515685.1:c931935-931327	135
NZ_KI515685.1:c931935-931327	346
NZ_KI515685.1:c931935-931327	405
NZ_KI515686.1:323579-324514	312
NZ_KI515686.1:323579-324514	453
NZ_KI515686.1:323579-324514	528
NZ_KI515686.1:323579-324514	596
NZ_KI515686.1:323579-324514	810
NZ_KI515686.1:613740-614315	105
NZ_KI515686.1:c200743-199319	457
NZ_KI515686.1:c200743-199319	742
NZ_KI515686.1:c642879-642748	108
NZ_KI515686.1:c642879-642748	94

CHAPTER IV

Discussion

Forensic genetics focuses on determining the source of biological evidence from a crime scene, most often by DNA profiling. However, biological evidence collected from crime scenes can have low quantities of DNA or be degraded and/or damaged (due to environmental or chemical factors). In some cases, there may not be enough human DNA in biological evidence to obtain a full, or even a partial, STR profile. Thus, there is a need for additional information to help identify the source of biological evidence found at crime scenes. The human skin microbiome is a potential source for targeted DNA analysis, as there is an abundance of microbes on the human skin.

The genetic content of the human microbiome likely exceeds that of the human body. The human microbiome is estimated to have approximately 232 million genes (1), compared to the approximately 45,000 (about 25,000 coding genes) genes annotated in the human genome. Touch DNA often refers to skin cells left behind by an individual touching or encountering an item. A touch DNA sample would also include the microorganisms that are shed with the skin cells. For every squamous cell shed, cells commonly found in the epidermis, there are approximately 30 microbial cells shed from the human skin (2). The high copy number and the high ratio of microbial cells compared to human skin cells suggests that the skin microbiome may provide more information about an individual than using transferred skin cells alone. Schmedes et al. (3) showed that samples collected directly from the skin could be used to extract human and microbial DNA. In (3), out of nine samples collected from an individual, only one sample produced a complete profile, while all nine samples produced enough microbial genetic information to accurately identify the individual they were collected from. Considering the samples were collected directly from the skin, lower amounts of human and microbial DNA would be recovered from touched items. The skin microbiome is in high abundance compared to human skin cells suggesting that trace human and microbial DNA profiles could be used together in source attribution of samples.

In this dissertation, the hypothesis that SNPs from stable, universal microbial species can differentiate skin microbiomes of individuals was investigated for its potential application towards forensic HID purposes. Schmedes et al. (3) developed a targeted clade-specific multiplex, called the hidSkinPlex, and described proof-of-concept work necessary for using the skin microbiome for HID. The studies within this dissertation focused on identifying informative SNPs contained within the markers of the hidSkinPlex that can be used for forensic HID. Samples taken from 51 individuals at three body-sites in triplicate were analyzed with the hidSkinPlex panel which targets specific skin microbiome organisms. SNPs with high estimates of Wright's fixation index (F_{ST}) were used in conjunction with supervised machine learning techniques (e.g., support vector machine (SVM) and least absolute shrinkage and selection operator (LASSO)) to select informative SNPs for the development of a new HID sequencing panel. Improved individualization was achieved by using informative SNPs and the associated allele frequencies in conjunction with supervised machine learning techniques to classify unknown samples.

In chapter 2, skin swabs from the non-dominant hand of 51 individuals were collected in triplicate and were analyzed for HID purposes. A predetermined number of the highest-ranking SNPs, based on their F_{ST} estimate, were selected using three different methods to determine if the number of taxa or SNPs impacted HID accuracies. The F_{ST} estimates for each SNP were then input into SVM to classify unknown samples to the individual they most resembled. Accuracy of classification ranged from 88% - 95%, suggesting that analysis of SNPs with high F_{ST} in targeted microorganisms can improve the accuracy of HID compared the abundance of taxa present in a sample (4, 5).

Chapter 3 presented an approach for determining a robust set of informative SNPs in the hidSkinPlex for HID. There were 51 individuals sampled on the ball of the foot, manubrium, and

non-dominant hand in triplicate that were sequenced using the hidSkinPlex panel. A mean F_{ST} estimate was calculated and used to rank all nucleotide positions in the data set. The nucleotide positions with a mean F_{ST} estimate > 0.1 and seen in at least 75% of sample comparisons were referred to as SNPs. LASSO was used to select of highly informative SNPs from the hidSkinPlex for attribution purposes. The final list of SNPs included 365 SNPs from 135 markers, specific to four species Cutibacterium acnes, Cutibacterium humerusii, Corynebacterium tuberculostearicum, and Propionibacteriaceae. These SNPs provided a 93% accuracy when identifying the host (n = 459). Based on these results a new panel, hidSkinPlex+, can be developed that may be more robust than the hidSkinPlex panel. This work supports that the skin microbiome may be a viable source for HID and potentially improve analysis of samples currently yielding low-level human nDNA, such as touch samples.

This dissertation provides some insight into how the hidSkinPlex might be further optimized. The hidSkinPlex is composed of 286 markers that range in size from 107 to 2,223 base pairs (bp), with a mean size of 601 ± 383 bp. Removing markers from the hidSkinPlex that do not contribute substantially to HID may improve the overall amplification efficiency and increase the sequencing coverage and as a result, more accurate information for the targeted region will be obtained. The hidSkinPlex+, a stream-lined version of hidSkinPlex, contains 365 SNPs selected in this study, and smaller sized amplicons, approximately <200 bp, could be designed such that approximately 150 primer pairs could capture all selected SNPs (Figure 1). Using amplicons that are 200 bp or less will allow for increased amplification efficiency. Some of the selected SNPs are contained within the same original marker from the hidSkinPlex and close together, allowing for the new amplicon size to capture more than one SNP that was selected for the hidSkinPlex+. A benefit of

designing primers to capture the most informative SNPs in a reduced amplicon size is that degraded DNA may be more readily analyzed.

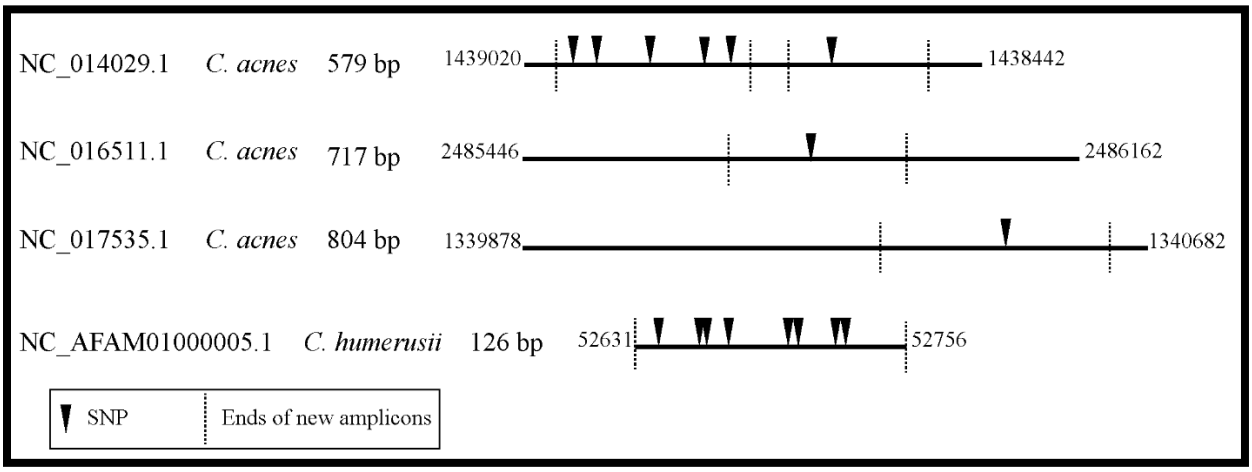


Figure 1. Four markers from hidSkinPlex are shown that could be redesigned into smaller amplicons for hidSkinPlex+. Each line contains the accession number, species, the original marker length from the hidSkinPlex panel, and then the marker is represented by a black line. The numbers flanking the line indicate nucleotide positions in the genome and SNPs from the hidSkinPlex+ are represented as triangles. Primers can be designed to capture the marked SNPs in smaller amplicons, where the vertical dashed lines indicate potential sites for primer design. While most markers from the hidSkinPlex will be reduced in length, some markers, such as the last marker highlighted at the bottom of the figure, will be kept the same size.

Additional research on the SNPs suggested for the hidSkinPlex+ is needed before the panel could be considered for use in a forensic setting. Population-level studies are needed so that the allele frequencies and covariance can be better estimated. Unlike traditional DNA markers, where human population groups are largely driven geography, the constitutive components of microbial population groups likely vary by geolocation, lifestyle, hygiene, and other environmental factors. Additionally, just as human DNA-based methods may be confounded by relatedness, microbial DNA-based methods may be equally confounded by close relationships. In particular, because microbial particles are routinely exchanged, relatedness or unrelatedness status may instead be not

just a function of heredity, but also of proximity. As such, future studies should address how genetics and environment influence the abundance and stability of the SNPs contained in the hidSkinPlex+ in different population groups with varying degrees of particle exchange (i.e., considering microbial "relatives"). Once the microbial allele frequency variation is better understood, alternative statistical analyses may be employed to associate a sample to an individual in a more traditional manner, such as with likelihood ratios.

In addition to larger sample sizes from multiple 'population' groups, other analysis methods for classification of an unknown sample should be investigated. SVM and logistic regression provided a strong foundation for classification of unknown data points. For example, both are linear methods that work well when the number of independent variables is larger than the sample size. However, linear methods are simplistic compared to non-linear machine learning algorithms. Non-linear methods may perform better for classification, but a fewer number of independent variables (e.g., SNPs) may be needed. Some appropriate nonlinear methods to consider include random forests, boosted tree-based methods, non-linear SVM, and convolution neural networks. The markers identified in the hidSkinPlex+ may allow these non-linear classifier algorithms to consider more nuanced relationships between SNPs within individuals and may in turn provide better HID.

Along with further optimization of the hidSkinPlex+, serious consideration will need to be given to the minimum requirements that must be met for a skin microbiome sample to be processed and then analyzed for inclusion of an individual as the possible donor. Minimum requirements for the input amount of microbial DNA needed for analysis, thresholds for read depth and minimum number of SNPs present in a profile should be determined. Employing thresholds will allow for associations that are made with a measure of confidence and provide a foundation for when and

how microbial profiling could be used in real casework. The time since deposit, collection, and processing could have an impact on the results derived from a skin microbiome sample. The hidSkinPlex+ contains markers that are already known to be common, abundant, and stable over time at multiple body sites. Testing the hidSkinPlex+ with time series samples will provide a better understanding of the stability of the selected SNPs and how F_{ST} may be able to handle when the SNPs have changed over time. Bringing together information about the stability of a sample after it is collected from a crime scene, the stability on an individual, and the minimum requirements for processing a skin microbiome sample is of the utmost importance if HID is going to be feasible using microorganism.

BIBLIOGRAPHY

1. Tierney BT, Yang Z, Luber JM, Beaudin M, Wibowo MC, Baek C, Mehlenbacher E, Patel CJ, Kostic AD. 2019. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. Cell Host & Microbe 26:283-295.e8.

2. Percival SL, Emanuel C, Cutting KF, Williams DW. 2012. Microbiology of the skin and the role of biofilms in infection. International Wound Journal 9:14-32.

3. Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. 2018. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Sci Int Genet 32:50-61.

4. Sherier AJ, Woerner AE, Budowle B. 2021. Population Informative Markers Selected Using Wright's Fixation Index and Machine Learning Improves Human Identification Using the Skin Microbiome. Appl Environ Microbiol 87:e0120821.

5. Woerner AE, Novroski NMM, Wendt FR, Ambers A, Wiley R, Schmedes SE, Budowle B. 2019. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci Int Genet 38:130-139.

BIBLIOGRAPHY

Adams RI, Bateman AC, Bik HM, Meadow JF. 2015. Microbiota of the indoor environment: a meta-analysis. Microbiome 3:49.

Alaeddini R, Walsh SJ, Abbas A. 2010. Forensic implications of genetic analyses from degraded DNA--a review. Forensic Sci Int Genet 4:148-57.

Banerjee AR. 2010. An Introduction to Conservation Genetics. The Yale Journal of Biology and Medicine 83:166-167.

Bogenhagen D, Clayton DA. 1974. The Number of Mitochondrial Deoxyribonucleic Acid Genomes in Mouse L and Human HeLa Cells: QUANTITATIVE ISOLATION OF MITOCHONDRIAL DEOXYRIBONUCLEIC ACID. Journal of Biological Chemistry 249:7991-7995.

Bosshard PP, Zbinden R, Abels S, Boddinghaus B, Altwegg M, Bottger EC. 2006. 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. J Clin Microbiol 44:1359-66.

Budowle B, Eisenberg AJ, van Daal A. 2009. Validity of low copy number typing and applications to forensic science. Croat Med J 50:207-17.

Butler JM. 2005. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, 2nd ed. Elsevier Academic Press, New York.

Clayton RA, Sutton G, Hinkle PS, Jr., Bult C, Fields C. 1995. Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. Int J Syst Bacteriol 45:595-9.

Doleckova I, Capova A, Machkova L, Moravcikova S, Maresova M, Velebny V. 2020. Seasonal variations in the skin parameters of Caucasian women from Central Europe. Skin Res Technol n/a.

Fay MP. 2010. Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. R Journal 2:53-58.

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 105:17994-9.

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 107:6477-81.

Fitz-Gibbon S, Tomida S, Chiu BH, Nguyen L, Du C, Liu M, Elashoff D, Erfe MC, Loncaric A, Kim J, Modlin RL, Miller JF, Sodergren E, Craft N, Weinstock GM, Li H. 2013. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. J Invest Dermatol 133:2152-60.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496-512.

Fox GE, Wisotzkey JD, Jurtshuk P, Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166-70.

Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930-8.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33:1--22.

Fujiyoshi S, Tanaka D, Maruyama F. 2017. Transmission of Airborne Bacteria across Built Environments and Its Measurement Standards: A Review. Front Microbiol 8:2336.

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 126:815-23.

Gorman B. 2018. Package 'mltools'. https://github.com/ben519/mltools. Accessed

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Program NCS, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190-2.

Grice EA, Kong HH, Renaud G, Young AC, Program NCS, Bouffard GG, Blakesley RW, Wolfsberg TG, Turner ML, Segre JA. 2008. A diversity profile of the human skin microbiota. Genome Res 18:1043-50.

Gu Y, Zha L, Yun L. 2017. Potential usefulness of SNP in the 16S rRNA gene serving as informative microbial marker for forensic attribution. Forensic Science International: Genetics Supplement Series 6:e451-e452.

Hampton-Marcell JT, Larsen P, Anton T, Cralle L, Sangwan N, Lax S, Gottel N, Salas-Garcia M, Young C, Duncan G, Lopez JV, Gilbert JA. 2020. Detecting personal microbiota signatures at artificial crime scenes. Forensic Sci Int 313:110351.

Hartl DL, Clark AG. 1997. Principles of Population Genetics. Sinauer Associates.

Heikens E, Fleer A, Paauw A, Florijn A, Fluit AC. 2005. Comparison of genotypic and phenotypic methods for species-level identification of clinical isolates of coagulase-negative staphylococci. J Clin Microbiol 43:2286-90.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-9.

Human Microbiome Jumpstart Reference Strains C, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT,

Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, et al. 2010. A catalog of reference genomes from the human microbiome. Science 328:994-9.

Human Microbiome Project C. 2012. A framework for human microbiome research. Nature 486:215-21.

Kapono CA, Morton JT, Bouslimani A, Melnik AV, Orlinsky K, Knaan TL, Garg N, Vazquez-Baeza Y, Protsyuk I, Janssen S, Zhu Q, Alexandrov T, Smarr L, Knight R, Dorrestein PC. 2018. Creating a 3D microbial and chemical snapshot of a human habitat. Sci Rep 8:3669.

Kapono CA, Morton JT, Bouslimani A, Melnik AV, Orlinsky K, Knaan TL, Garg N, Vazquez-Baeza Y, Protsyuk I, Janssen S, Zhu Q, Alexandrov T, Smarr L, Knight R, Dorrestein PC. 2018. Creating a 3D microbial and chemical snapshot of a human habitat. Sci Rep 8:3669.

Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR. 2014. Progress toward an efficient panel of SNPs for ancestry inference. Forensic Sci Int Genet 10:23-32.

Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328-33.

Knight R, Metcalf JL, Gilbert JA, Carter DO. 2018. Evaluating the Skin Microbiome as Trace Evidence. National Criminal Justice Reference Service.

Kwong JC, McCallum N, Sintchenko V, Howden BP. 2015. Whole genome sequencing in clinical and public health microbiology. Pathology 47:199-210.

Lax S NC, Gilbert JA. 2015. Our interface with the built environment: immunity and the indoor microbiota. Trends Immunol.

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

Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S, Metcalf JL, Ursell LK, Vazquez-Baeza Y, Van Treuren W, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, Gilbert JA. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science 345:1048-52.

Lee S-Y, Woo S-K, Lee S-M, Eom Y-B. 2016. Forensic analysis using microbial community between skin bacteria and fabrics. Toxicology and Environmental Health Sciences 8:263-270.

Leung MHY, Tong X, Wilkins D, Cheung HHL, Lee PKH. 2018. Individual and household attributes influence the dynamics of the personal skin microbiota and its association network. Microbiome 6:26.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.

Luongo JC, Barberán A, Hacker-Cary R, Morgan EE, Miller SL, Fierer N. 2017. Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. Indoor Air 27:338-344.

Lynch M. 2003. God's signature: DNA profiling, the new gold standard in forensic science. Endeavour 27:93-7.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:3.

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

Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, Brown GZ, Green JL, Bohannan BJ. 2014. Bacterial communities on classroom surfaces vary with human contact. Microbiome 2:7.

Meyer D, Evgenia Dimitriadou, Hornik K, Weingessel A, Leisch F. 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3. https://CRAN.R-project.org/package=e1071.

Mignard S, Flandrois JP. 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. J Microbiol Methods 67:574-81.

Neckovic A, van Oorschot RAH, Szkuta B, Durdle A. 2020. Investigation of direct and indirect transfer of microbiomes between individuals. Forensic Sci Int Genet 45:102212.

Oh J, Byrd AL, Deming C, Conlan S, Program NCS, 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, Program NCS, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. Cell 165:854-66.

Park J, Kim SJ, Lee J-A, Kim JW, Kim SB. 2017. Microbial forensic analysis of human-associated bacteria inhabiting hand surface. Forensic Science International: Genetics Supplement Series 6:e510-e512.

Percival SL, Emanuel C, Cutting KF, Williams DW. 2012. Microbiology of the skin and the role of biofilms in infection. International Wound Journal 9:14-32.

Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Borsting C, Johansen P, Fondevila M, Morling N, Schneider P, Consortium EU-N, Carracedo A, Lareu

MV. 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic Sci Int Genet 11:13-25.

Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, van Schaik W, Wertheim HFL. 2017. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev 30:1015-1063.

Richardson M, Gottel N, Gilbert JA, Lax S. 2019. Microbial Similarity between Students in a Common Dormitory Environment Reveals the Forensic Potential of Individual Microbial Signatures. mBio 10:e01054-19.

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

Schmedes SE, Woerner AE, Budowle B. 2017. Forensic Human Identification Using Skin Microbiomes. Appl Environ Microbiol 83.

Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. 2018. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Sci Int Genet 32:50-61.

Sender R, Fuchs S, Milo R. 2016. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell 164:337-40.

Sherier AJ, Woerner AE, Budowle B. 2021. Population Informative Markers Selected Using Wright's Fixation Index and Machine Learning Improves Human Identification Using the Skin Microbiome. Appl Environ Microbiol 87:e0120821.

Song SJ, Lauber C, Costello EK, Lozupone CA, 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.

Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625-30.

Tang YW, Ellis NM, Hopkins MK, Smith DH, Dodge DE, Persing DH. 1998. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. J Clin Microbiol 36:3674-9.

Team RC. 2013. R: A Language and Environment for Statistical Computing, *on* R Foundation for Statistical Computing. http://www.R-project.org/. Accessed 2/10/21.

The Human Microbiome Project C, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207.

Tierney BT, Yang Z, Luber JM, Beaudin M, Wibowo MC, Baek C, Mehlenbacher E, Patel CJ, Kostic AD. 2019. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. Cell Host & Microbe 26:283-295.e8.

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 12:902-3.

Wang Y, Yu Q, Zhou R, Feng T, Hilal MG, Li H. 2021. Nationality and body location alter human skin microbiome. Applied Microbiology and Biotechnology doi:10.1007/s00253-021-11387-8.

Watanabe H, Nakamura I, Mizutani S, Kurokawa Y, Mori H, Kurokawa K, Yamada T. 2018. Minor taxa in human skin microbiome contribute to the personal identification. PLoS One 13:e0199947.

Wickham H, Averick M, Bryan J, Chang W, McGowan LDAF, Romain , Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the {tidyverse}. Journal of Open Source Software 4:1686.

Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K. 2016. ggplot2: Elegant Graphics for Data Analysis, vol 2018. Springer-Verlag New York.

Woerner AE, Novroski NMM, Wendt FR, Ambers A, Wiley R, Schmedes SE, Budowle B. 2019. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci Int Genet 38:130-139.

Woo PC, Ng KH, Lau SK, Yip KT, Fung AM, Leung KW, Tam DM, Que TL, Yuen KY. 2003. Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. J Clin Microbiol 41:1996-2001.

Wright S. 1949. The Genetical Structure of Populations. Annals of Eugenics 15:323-354.

Wright S. 1951. The genetical structure of populations. Ann Eugen 15:323-54.

Yang J, Tsukimi T, Yoshikawa M, Suzuki K, Takeda T, Tomita M, Fukuda S. 2019. Cutibacterium acnes (Propionibacterium acnes) 16S rRNA Genotyping of Microbial Samples from Possessions Contributes to Owner Identification. mSystems 4:e00594-19.

Zeng X, Chakraborty R, King JL, LaRue B, Moura-Neto RS, Budowle B. 2016. Selection of highly informative SNP markers for population affiliation of major US populations. Int J Legal Med 130:341-52.