



HOKKAIDO UNIVERSITY

Title	Studies on identification and evaluation of CRISPR diversity on human skin microbiome for development of a new personal identification method
Author(s)	豊間根, 耕地
Degree Grantor	北海道大学
Degree Name	博士(感染症学)
Dissertation Number	乙第7170号
Issue Date	2022-12-26
DOI	https://doi.org/10.14943/doctoral.r7170
Doc URL	https://hdl.handle.net/2115/88079
Type	doctoral thesis
File Information	Kochi_Toyomane.pdf



**Studies on identification and evaluation of CRISPR
diversity on human skin microbiome for development of a
new personal identification method**

(新規個人識別法の開発を目的としたヒト皮膚マイクロバイオーームにおける CRISPR の探索と多様性解析)

Kochi Toyomane

NOTES

The contents of Chapter I and Chapter II have been published in mSystems.

Toyomane K, Yokota R, Watanabe K, Akutsu T, Asahi A, and Kubota S. (2021)

Evaluation of CRISPR Diversity in the Human Skin Microbiome for Personal Identification. mSystems **6**(1). DOI: 10.1128/mSystems.01255-20

© 2021 Toyomane et al. CC BY 4.0

The contents of Chapter III have been published in Forensic Science International: Reports.

Toyomane K, Yokota R, Watanabe K, Akutsu T, Asahi A, and Kubota S. (2022)

Optimization of microbial DNA extraction from human skin samples for CRISPR typing. Forensic Science International: Reports **5**. DOI: 10.1016/j.fsir.2022.100259

© 2022 Toyomane et al. CC BY-NC-ND 4.0

CONTENTS

NOTES	i
CONTENTS	ii
ABBREVIATIONS	1
PREFACE	1
CHAPTER I Metagenomic reconstruction of CRISPR arrays in the human skin microbiome	6
Introduction	6
Materials and methods	7
Results	9
Discussion	13
Summary	17
CHAPTER II Evaluation of CRISPR diversity in the human skin microbiome for personal identification	27
Introduction	27
Materials and methods	28
Results	35
Discussion	41
Summary	46
CHAPTER III Optimization of microbial DNA extraction from human skin samples for CRISPR typing	59
Introduction	59
Materials and methods	61

Results	64
Discussion	66
Summary	68
CONCLUSION	73
ACKNOWLEDGMENTS.....	75
和文要旨	76
REFERENCE	79

ABBREVIATIONS

Ac	antecubital fossa
ASV	amplicon sequence variant
Cas	CRISPR-associated protein
CRISPR	clustered regularly interspaced short palindromic repeat
EZ1	EZ1 DNA Investigator Kit
Hp	hypothenar palm
IQR	interquartile range
OTU	operational taxonomic unit
PCoA	Principal coordinates analysis
PS	DNeasy PowerSoil Kit
Ra	retroauricular crease
SRA	Sequence Read Archive
STR	short tandem repeat
SVM-RFE	Support Vector Machine based on Recursive Feature Elimination

PREFACE

Microbiome is defined as “the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions,” based on the concept of “biome,” which contains both biological members and non-biological factors in a given environment (1). Microbiome is ubiquitous in but unique to each ecosystem; while every ecosystem including humans, animals and even office environments harbors microorganisms, the taxonomic composition, functional interactions and temporal dynamics of the microbiome are distinct each other (2-4). Furthermore, it has been revealed that human microbiome, the microbial communities on human body, are related to host health and diseases, such as type 2 diabetes, atopic dermatitis, and pre-term birth, attracting increasing attention in recent biomedical researches (5-7).

While original researches utilized culture-based approach to characterize the microbiome, non-culturable microbes are not detected by culture process. Culture-independent approach which utilize sequencing technology emerged in order to capture both culturable and non-culturable microbes (8). Currently, two major methodologies, amplicon sequencing and shotgun metagenome sequencing, are used (9). Amplicon sequencing is a technique which uses primers targeting a taxonomically informative gene to amplify the gene by PCR. Usually, the 16S rRNA or the internal transcribed spacer region are sequenced for bacterial or fungal communities, respectively. The PCR amplicon are then analyzed by a sequencer to find relative abundances of the microbial taxa at genus-level and species-level. Shotgun metagenome sequencing, on the other hand, analyzes all genomic contents within a sample, which have enabled strain-level profiling

of the microbiome. In addition, it is also possible to capture catalogue of functional genes by using shotgun metagenome sequencing, while it is relatively expensive and laborious (10).

Rapid advances in sequencing technology have drastically changed the microbiome researches, which have revealed significant variation in the structure of the human skin microbiome among healthy individuals (11-13). Fierer et al. demonstrated that bacterial communities on the hands of different people share only 13% identity (14). Several studies have demonstrated individuality of the microbiome as a “bacterial fingerprint” to develop a novel method for the identification of people involved in crime scenes, in which metagenome datasets or 16S rRNA sequencing data were analyzed (11, 15, 16). It is expected that bacterial DNA can be easily recovered from the surface of objects touched by a suspect, considering the abundance of bacteria on human skin, which is estimated to be approximately 10000 cells/cm² (11, 17). The microbiome of the human skin has gained attention in recent years owing to low success rates of DNA profiling from an object touched by hands, which typically yields a small amount of DNA (18). Considering the number of samples for DNA profiling in Japan, which reaches approximately 300 thousand samples per year, it will be beneficial to achieve effective and reliable police activities that development of a new profiling method which can complement conventional DNA profiling by handling low human DNA samples (19).

Identification of individuals in crime scenes based on microbial community diversity requires characterization of the microbiome community with a degree of high resolution. Schmedes et al. utilized whole-genome shotgun sequencing to discriminate personal microbiomes using a strain-level taxonomic profiling (15). Although the study achieved high accuracy for personal identification by analyzing public data, the high cost

and complexity of analysis precludes its practical use. In contrast, amplicon-based microbiome profiling is comparatively cost-effective as it can be performed using conventional methods such as microarray analysis (20-23). As a phylogenetic marker, 16S rRNA gene has been widely used in microbiome studies. However, 16S rRNA-based personal identification is easily perturbed over time (24, 25). The temporal instability of the “bacterial fingerprint” is likely due to loss of minor species in an environment. Moreover, the resolution of amplicon sequencing using 16S rRNA is not sufficient to discriminate bacteria at the strain level, or, in many cases, at the species-level (26). The importance of strain-level classification of specific species for personal identification has been previously proposed (15, 27). Effective personal identification would require determination of the microbial community composition at high resolution; thus, novel strategies are needed to employ amplicon sequencing for forensic purposes.

The clustered regularly interspaced short palindromic repeat (CRISPR) array corresponds to a genomic locus found in prokaryotes and is composed of identical repeats, formerly called direct repeats, and unique sequences between repeats, called spacers (28). Typically, CRISPR repeats range from 28 to 37 bp in length, whereas spacers range between 32 and 38 bp (29). Each CRISPR corresponds roughly to a specific taxon. However, some strains harbor multiple CRISPR loci and some CRISPRs are conserved in multiple species. CRISPRs along with CRISPR-associated proteins (Cas), comprise the adaptive immune system of prokaryotes (30). It serves as a defensive mechanism against foreign genetic entities, such as phages. To provide immunity against the re-infecting foreign DNA, the CRISPR-Cas system integrates a short DNA fragment derived from the foreign DNA as a new spacer. Therefore, spacer sequences in a CRISPR array represent a history of past infection encountered by a bacterial strain and reflect the

structure of the viral community, which is more personalized than that of the bacterial community itself (31, 32).

The present thesis aimed to develop and evaluate a new amplicon sequencing method which analyzes personalized skin microbiome with resolution of strain-level for forensic personal identification as a proof-of-concept. To achieve strain-level classification of skin microbiome, analytical method to evaluate diversity of CRISPR sequence in skin microbiome were developed. In chapter I, the diversity of CRISPR is evaluated to identify the most abundant CRISPR repeats in the human skin microbiome by mining publicly available datasets. In chapter II, CRISPR typing by the amplicon sequencing methodology was developed to analyze spacer diversity of three candidate CRISPRs in the skin microbiome. Then, the data was also used to predict the origin of the samples by a machine learning approach to evaluate the potential of CRISPR typing as a tool for personal identification. In chapter III, two DNA extraction protocol were compared to optimize the protocol of the metagenomic CRISPR typing developed in chapter II for forensic purpose. In addition, longitudinal stability of the CRISPR diversity was also evaluated by expanding the results in chapter II with results of samples newly analyzed in chapter III.

CHAPTER I

Metagenomic reconstruction of CRISPR arrays in the human skin microbiome

Introduction

As variable spacers in a CRISPR array are flanked by conserved repeats, it can be used as a phylogenetic marker that provides higher resolution than that of other markers, such as 16S rRNA. CRISPR typing, which is also called spacer oligonucleotide typing (spoligotyping), using the spacer acquisition history of CRISPR arrays has been utilized for phylogenetic analysis of bacteria (33-42, and see reference 43 for a review). This typing strategy was also applied in some metagenomic studies to characterize complex microbial populations or identify host-virus interactions, both by reconstruction of CRISPRs from shotgun metagenomic datasets (44-47) and amplicon sequencing utilizing primers targeting the repeats of CRISPR (32, 44, 48-50). A limitation of CRISPR typing for microbiome characterization is that the spacers of a specific CRISPR locus recovered from a shotgun metagenome dataset may be insufficient to understand spacer diversity within the microbial community. For example, Rho et al. could not identify the distribution of CRISPR arrays in some skin datasets by CRISPR array reconstruction using shotgun metagenome datasets obtained from the Human Microbiome Project (45). Another limitation is that amplicon sequencing targeting CRISPR arrays is strongly dependent on *a priori* knowledge of repeat sequences (44); CRISPR spacers cannot be sequenced by amplicon sequencing without understanding which CRISPR arrays are in a specific environment, whereas amplicon sequencing can be utilized for complete

sequencing of spacers in a specific CRISPR locus. Therefore, it remains unclear whether it is possible to discriminate personal microbiomes using spacer diversity or identify the CRISPRs that are suitable for personal identification. To overcome these limitations, the identity of CRISPRs that are common in the human skin microbiome must be determined to enable characterization of the spacer diversity of CRISPRs in different individuals.

In this chapter, the most abundant CRISPR repeats in the human skin microbiome were identified via metagenomic reconstruction of CRISPR arrays. In addition, number of reads containing CRISPR repeats were determined by an assembly-independent manner, in order to identify the frequencies with which repeats are shared within the individuals.

Materials and methods

Metagenomic datasets

Metagenomic datasets from the skin of the palm were downloaded from the European Nucleotide Archive in fastq format. The samples used in this study have been registered under BioProject PRJNA46333. Data collection methods were performed in accordance with previously reported method by Oh et al. (12). Samples containing multiple runs were merged before assembly.

Identification of CRISPR arrays using metagenomic datasets and isolation of CRISPRs conserved across multiple individuals

The scheme for CRISPR reconstruction from the metagenomic reads and isolation has been summarized in Figure 1. Reads were filtered using Trimmomatic 0.39 to remove low-quality reads (51), bases with low-quality scores ($Q < 20$) were trimmed

from both ends, and short reads (< 50 bp) were removed. As the file was too large to assemble, 8×10^7 reads (22%) were subsampled from the largest dataset (SRX743731) using seqtk (<https://github.com/lh3/seqtk>). De novo assembly of the metagenome was performed using MEGAHIT v1.1.4 (52). The datasets were assembled with default parameters, except that the k-mer sizes were fixed as --k-list 29,39,59,79,99. Output contigs were analyzed using the PILER-CR v1.06 with default parameters to detect the CRISPR arrays (53). Each repeat in a CRISPR array detected by the PILER-CR was converted to a single read. All reads were then duplicated and one of the read pairs was reverse complemented, as the orientation of repeat sequence was unclear.

Repeat reads were then analyzed using QIIME 2 v2019.7 (54). First, dereplication and de novo operational taxonomic unit (OTU) clustering at 90% were performed using q2-vsearch plug-in (55). The repeat reads were detected in less than two individuals and the duplicated reverse complement repeat reads were discarded. The representative CRISPR repeats of each OTU were annotated by BLAST using default settings (56). The subject with the highest max score was defined as the host species. Since CRISPRs are mobile genetic elements, multiple species were often found to share the same repeat (57); in such cases, one representative species was recorded. Eventually, heatmaps of the repeat reads on the OTU table were constructed using the heatmap visualizer of the q2-feature-table plug-in (58). The type and orientation of CRISPRs were determined by a search for same repeats in CRISPRCasdb (59).

CRISPRCasFinder was utilized to identify *cas* genes associated with the CRISPR locus found in metagenomic contents (60). The contigs assembled by MEGAHIT were used as initial inputs. CRISPRCasFinder was then used with default settings in order to extract the contigs that contained CRISPRs with evidence-level 4 and

were > 1000 bp in length. The evidence-level was determined by CRISPRCasFinder. Subsequently, the extracted contigs were subjected to CRISPRCasFinder again to determine if the contigs contain *cas* genes and to subtype the identified *cas* genes based on the nomenclature proposed by Makarova et al. (61) using CRISPRCasFinder with the -cas parameter. Three representative contigs containing STEQ CRISPR arrays were compared with the genome of *Streptococcus* strains (Table 1) using CRISPR Visualizer (62).

Measurement of repeats in the metagenomic dataset by BLAST search

The number of reads containing each repeat in a dataset was measured using BLAST. The sequences of repeats were used as queries and the skin metagenome datasets in the Sequence Read Archive (SRA) of NCBI were used as subjects. The datasets used as subjects were identical to those used in the CRISPR reconstruction, except for samples from nares or cheek. Two samples (SRX740760 and SRX740761) from nares or cheek, representing oily and moist sites, respectively, of the same individual (HV07) were also included to evaluate repeat variation in different skin sites within an individual. Default parameters were used for the short-read search except that the maximum number of target sequences was set as 5000. Reads with > 90% identity and > 90% query cover were included in the downstream analysis. A hierarchically clustered gradient heat map of repeat frequency was plotted using the clustermap method in the Seaborn package 0.9.0 (<https://seaborn.pydata.org/generated/seaborn.clustermap.html>) in Python 3.6.8.

Results

Identification of conserved CRISPR arrays among individuals in skin using

metagenome datasets

Amplicon-based CRISPR sequencing cannot be used without prior knowledge of the CRISPRs present in a microbial community. Thus, the diversity of CRISPRs in the human skin microbiome was investigated by analyzing metagenome datasets obtained in a previous study (12) via metagenomic reconstruction of CRISPRs. Although reconstructions of the CRISPR arrays in the human microbiome were previously reported (45, 46), the diversity of CRISPR arrays in the skin metagenomic data remained unclear. In this study, the data of samples collected from the Hp region were analyzed, as DNA from the hands are most likely to be collected from crime scenes. The metagenomic data were assembled using the de novo assembler MEGAHIT (52). CRISPR arrays were identified by PILER-CR, a tool developed for identifying CRISPRs (53). Using this pipeline, 862 arrays were identified from 15 samples among 14 individuals (Table 2). No CRISPR arrays were obtained from samples MET0022 (HV01), MET0161 (HV08), or MET0177 (HV07). This may be attributed to insufficient assembly, as the number of reconstructed CRISPR arrays greatly depends on the assembly quality. Interestingly, two samples, MET0177 (HV07; Hp-L) and MET0190 (HV07; Hp-R) derived from the same individual, yielded metagenomes with different detectable CRISPR arrays. These results suggest that the repeatability of metagenomic reconstruction is insufficient to fully measure CRISPR diversity.

The origins of the repeats detected via metagenomic CRISPR reconstruction were identified by a BLAST search. Among 24 CRISPR repeat sequences, 14 CRISPR repeats were identical to their respective reference genomes, while the other 10 did not match exactly to the reference genomes (Table 3). Thus, the latter repeats may have been derived from previously unknown CRISPR arrays, although it remains unknown whether

these putative CRISPRs were correctly reconstructed. Note that, in such cases, the annotation of CRISPRs including host species, subtyping, or orientation might not be precise or be impossible and should be clarified in the future studies. The most frequently observed repeat was derived from *Streptococcus equinus* (STEQ), which was observed in eight individuals (Figure 2A, Table 2). This specific streptococcal repeat has been targeted in other studies using specific primers to amplify the spacers of *Streptococcus* species in the oral and skin microbiomes (50). In total, 24 putative CRISPR arrays were observed for at least two individuals, indicating that some CRISPR arrays were conserved among the population. To validate the reconstructed CRISPR arrays, the CRISPR-Cas systems were subtyped based on nomenclature proposed by Makarova et al. using CRISPRCasFinder and compared the arrays with a known CRISPR database if *cas* genes were also detectable (60, 61). Using CRISPRCasFinder, 42 CRISPRs that associated with *cas* genes were identified (Table 4); a representative CRISPR-Cas locus is shown in Figure 2C. Two CRISPR arrays were compared with the most relevant arrays in a public database. For the most abundant CRISPR, STEQ, three CRISPR arrays were detected by CRISPRCasFinder, one of which was subjected to a BLAST search to identify the most relevant CRISPR loci in the NCBI database. It was found that the most similar CRISPR-Cas region for STEQ CRISPR was that of *Streptococcus oralis* strain FDAARGOS_367 (GenBank: CP023507) and STEQ CRISPR was classified into type IIA, which is characterized by *cas9* gene. Another example of CRISPR loci with *cas* genes was that of *Streptococcus thermophilus* (STTH). For STTH CRISPR, one contig containing both CRISPR and *cas* genes was obtained. This CRISPR-Cas was close to that of *Streptococcus mitis* strain S022-V3-A4 (GenBank: CP047883), which was also classified into type IIA. Repeats of both STEQ CRISPR and STTH CRISPR were identical to those

of CRISPR1 and CRISPR3 of *Streptococcus thermophilus* (63), although the entire loci were close to other *Streptococcal* species, suggesting that these CRISPRs are widely preserved among *Streptococcus*. Then, three representative STEQ CRISPRs, which were detected with *cas* genes, were compared against those with the same repeats (Figure 3). Given that all spacers in the reconstructed arrays were unique among the 25 references in CRISPRCasdb (59), it is difficult to precisely identify the array origin. Expansion of the CRISPR database may improve classification accuracy in the future. Altogether, the results suggest that CRISPR reconstruction using metagenomic data accurately detects repeats present in the skin microbiome.

CRISPR repeat abundance in the skin microbiome

Insufficient CRISPR arrays were detected in multiple datasets, with some individuals sharing no CRISPR arrays with others, perhaps due to the low coverage of CRISPR arrays. Reads containing CRISPR repeats may be observed more frequently in metagenome datasets, even in those where no CRISPR arrays were reconstructed. For repeats observed in metagenomic CRISPR reconstruction, the number of reads containing each repeat were measured by using BLAST. The reads containing repeats were enumerated by searching the repeat sequence as a query against the metagenome dataset using the SRA as a subject. Repeats observed in at least three individuals in the CRISPR reconstructions were included as queries for read counting. Among the CRISPR repeats identified in CRISPR reconstructions, the repeats derived from STEQ, STTH, *Neisseria polysaccharea*, and *Syntrophomonas wolfei* were most frequently observed in read counting (Figure 2B, Table 5; repeat identifiers are listed in Table 2). Reads containing the repeats of these CRISPRs were detected from the hands of 14 of the 15 individuals

by read counting, while those CRISPRs were observed in 8, 5, 6, and 3 individuals, respectively, by metagenomic CRISPR reconstructions (Table 3). As shown in Figure 2B, individuals were categorized into three clusters based on the most abundant repeat; the most abundant repeats in the first and the second groups were derived from *Micrococcus luteus* (MILU) and *Chryseobacterium*, respectively. In the last group, repeats derived from MILU and *Chryseobacterium* were less abundant than other repeats, such as those derived from STEQ or *N. polysaccharea*. Clustering of repeat abundances was plausible because four samples from the same individual (HV07), including two samples each from nares and cheek (representing moist and oily sites, respectively), were grouped into the same cluster. Some repeats were observed in individuals even when no CRISPR array was detected in the assembled contigs (HV01 and HV08), suggesting that the prevalence of CRISPR arrays was underestimated by metagenomic reconstruction. Therefore, amplicon sequencing may be a more effective tool for studying CRISPR diversity.

Discussion

Regardless of previous studies that utilized amplicon sequencing for evaluation of spacer diversity in oral, skin, or lake environments (32, 44, 48-50), the high inter-individual diversity of spacer sequences on human skin is still largely unknown, given that the specific CRISPR arrays in skin environments commonly shared among individuals have not been clarified. This study identified putative CRISPRs shared between the skin microbiomes of different individuals by reconstructing CRISPR arrays using publicly available metagenome datasets (Figure 2).

Metagenomic reconstruction of CRISPR arrays identified the consensus repeats of CRISPR arrays conserved among individuals. The repeats of CRISPRs derived from

STEQ, *N. polysaccharea*, STTH, *V. parvula*, *Aggregatibacter aphrophilus*, and *S. wolfei* were observed in 14 of 15 individuals (Figure 2B), suggesting that CRISPR arrays with those repeats are widely spread among the human population. This observation is partially in agreement with a study by Robles-Sikisaka et al. (50), in which spacer diversity of *Streptococcus*-derived CRISPRs in the human skin microbiome was investigated by amplifying the spacer regions with primers targeting the repeat derived from STEQ.

Interestingly, CRISPR arrays derived from *Cutibacterium acnes*, which is the most abundant microorganism in the skin microbiota (12, 13), were rarely detected in the current study. This discrepancy may be attributed to the scarcity of CRISPR arrays in populations of *C. acnes*, as only one lineage of *C. acnes* among the three major groups harbors CRISPR arrays and those in the organism harbor a limited number of spacers (1–9 spacers, with a mean of 3.31) (64–66). Fitz-Gibbon et al. reported that only two of the top 10 most abundant *C. acnes* ribotypes are associated with CRISPR-Cas and that the CRISPR-encoding ribotypes are not major in > 80 % of individuals (65), supporting the limited observation of the *C. acne* CRISPRs. In contrast, the repeats detected in the metagenomic CRISPR reconstruction were derived from minor genera in the skin metagenomic data, such as *Chryseobacterium* and *Aggregatibacter*, which comprised <0.10222% of reads in the metagenomic dataset produced by Oh et al. (12). The present study is the first to evaluate the ecology of CRISPRs in the skin microbiome, identifying previously unknown CRISPRs and evaluating CRISPR abundance in the human-associated microbiome, providing insight into the nature of the skin microbiome.

CRISPR arrays metagenomically reconstructed were also compared with reference arrays by CRISPR Visualizer. It is shown that all spacers in reconstructed CRISPRs are unique among 25 reference strains, suggesting that CRISPRs in skin

microbiome are distinct from each other. This high diversity is in agreement with the previous reports demonstrating highly personal specific nature of virus community in the human microbiome, because spacers of a CRISPR are derived from viral genome (12, 31, 32). Thus, diversity of CRISPRs may be utilized as a “bacterial fingerprint”, while it is required to comprehensively characterize CRISPR diversity in skin microbiome by amplicon sequencing.

While metagenome datasets were analyzed by assembly-based approach to analyze CRISPRs in this chapter, some issues are still remained. First, repeats were found in the samples in which no CRISPR were obtained by assembly-based approach, as mentioned above. In general, repeat region causes gaps between contigs (67). As CRISPRs are repetitive elements, it is not surprising if the assembly errors are induced at CRISPR loci, which leading to failure in PILER-CR, a program based on repeat detection algorithm and used to identify CRISPRs from genome sequences. The failure of assembly at CRISPR locus was also supported by the representative CRISPR which is shown in Figure 2 because this CRISPR locus was also truncated due to assembly failure. Second, while reconstruction of CRISPR from entire metagenome can identify previously unknown CRISPRs, the number of reads derived from a specific CRISPR may be insufficient to reconstruct whole locus. Indeed, the number of reads containing a specific repeat was 4994 at most (Table 5), supporting the limitation of the metagenomic approach. Altogether, it suggested that an amplicon sequencing methodology should be developed to explore high diversity of a CRISPR in an environment.

In this chapter, 24 putative CRISPRs were identified by reanalysis of publicly available metagenomic datasets. The repeat sequences of CRISPR arrays were found in up to 14/15 individuals, indicating that CRISPR arrays with those repeats are widely

spread among the human population. Furthermore, all metagenomically reconstructed spacers are unique among the 25 reference *Streptococcus* CRISPRs, suggesting high diversity of streptococcal CRISPRs. In conclusion, these data indicated the potential of CRISPRs as a new marker for characterization of personalized human skin microbiome.

Summary

The highly personalized human skin microbiome may serve as a viable marker in personal identification. Amplicon sequencing resolution using 16S rRNA cannot identify bacterial communities sufficiently to discriminate between individuals. Thus, novel, higher resolution genetic markers are required for forensic purposes. CRISPRs are prokaryotic genetic elements that can provide a history of infections encountered by the bacteria. The sequencing of CRISPR spacers may provide phylogenetic information with higher resolution than other markers. However, using spacer sequencing for discrimination of personal skin microbiome is difficult due to limited information on CRISPRs in human skin microbiomes. It remains unclear whether personal microbiome discrimination can be achieved using spacer diversity, or which CRISPRs will be forensically relevant. Metagenomic reconstruction was utilized to identify common CRISPRs in the human skin microbiome in order to identify CRISPRs which can be used as personal identification marker. As a result, 24 putative CRISPR arrays were successfully reconstructed using metagenomic datasets. Abundance of putative CRISPRs were confirmed by measurement of CRISPR repeats in a metagenomic assembly-independent manner. In addition, comparison of CRISPRs identified by metagenomic reconstruction revealed uniqueness of CRISPR arrays found in skin microbiome, suggesting that CRISPRs identified in this chapter would be useful marker for the characterization of personal microbiome.

Table 1 List of Streptococcus strains used as reference CRISPRs.

Strain	Accession #	Note
Streptococcus agalactiae 2603V:R	NC_004116.1	
Streptococcus agalactiae A909	NC_007432.1	
Streptococcus agalactiae GD201008-001	CP003810.1	
Streptococcus agalactiae NEM316	NC_004368.1	
Streptococcus lutetiensis 033	NC_021900.1	Two CRISPRs sharing same repeats were identified by CRISPRviz
Streptococcus mutans GS-5	CP003686.1	
Streptococcus mutans LJ23	NC_017768.1	
Streptococcus mutans NN2025	NC_013928.1	
Streptococcus mutans UA159-FR	NZ_CP007016.1	
Streptococcus mutans UA159	NC_004350.2	
Streptococcus mutans strain NG8	CP013237.1	
Streptococcus thermophilus ASCC 1275	NZ_CP006819.1	
Streptococcus thermophilus JIM 8232	NC_017581.1	
Streptococcus thermophilus LMD-9	NC_008532.1	
Streptococcus thermophilus MN-ZLW-002	NC_017927.1	
Streptococcus thermophilus ND03	NC_017563.1	
Streptococcus thermophilus strain CS18	NZ_CP030928.1	
Streptococcus thermophilus strain CS5	NZ_CP028896.1	
Streptococcus thermophilus strain CS9	NZ_CP030927.1	
Streptococcus thermophilus strain DGCC 7710	CP025216.1	
Streptococcus thermophilus strain KLDS 3.1003	NZ_CP016877.1	
Streptococcus thermophilus strain KLDS SM	NZ_CP016026.1	
Streptococcus thermophilus strain MN-BM-A01	NZ_CP012588.1	
Streptococcus thermophilus strain MN-BM-A02	NZ_CP010999.1	
Streptococcus thermophilus strain SMQ-301	NZ_CP011217.1	

Table 2 Summary of metagenomic CRISPR reconstruction using MEGAHIT and PILER-CR.

Experiment	SAMPLE	Patient ID	Gender	Age range	Site-Symmetry	Total initial reads pair	Too short after quality trimming (<50bp)	Forward only	reverse only	Read pair trimming discrepancy	Total final reads pair	% surviving reads	Reads used for assembly	Contigs	Total length (bp)	Minimum contig length (bp)	Max contig length (bp)	Average contig length (bp)	N50 (bp)	CRISPR arrays
SRX739783	MET0022	HV01	Male	30s	Hp-R	2195791	44075	156983	45383	202366	1949350	88.78	3898700	3879	4319165	203	152685	1113	14651	0
SRX740541	MET0041	HV03	Female	30s	Hp-R	8650474	228189	814318	288853	1103171	7319114	84.61	14638228	49111	38948159	200	50006	793	976	41
SRX740562	MET0065	HV02	Female	30s	Hp-R	2353252	118865	324146	125893	450039	1784348	75.82	3568696	10732	5486321	203	4300	511	499	1
SRX740589	MET0094	HV05	Male	20s	Hp-R	5993912	226267	934709	256667	1191376	4576269	76.35	9152538	37854	26519835	200	25008	701	761	20
SRX740632	MET0139	HV06	Female	30s	Hp-R	11836038	213642	525767	123789	649556	10972840	92.71	21945680	97822	61019513	200	73932	624	645	39
SRX740643	MET0161	HV08	Male	30s	Hp-L	3375857	30652	73693	32244	105937	3239268	95.95	6478536	18927	10797241	202	24453	570	524	0
SRX740842	MET0177	HV07	Male	20s	Hp-L	19702061	25852	20636	1761	22397	19653812	99.76	39307624	15546	9391979	206	6144	604	685	0
SRX741396	MET0190	HV07	Male	20s	Hp-R	87777711	25256	5686	670	6356	8746099	99.64	17492198	65733	55065694	200	355222	838	999	18
SRX743731	MET0207	HV09	Male	30s	Hp-R	362229982	2099578	253204	32826	32826	359844374	99.34	80000000	84118	36044104	201	18239	428	391	20
SRX743869	MET0224	HV10	Male	30s	Hp-R	28464770	34601	22773	1907	24680	28405489	99.79	56810978	120680	122476808	200	230363	1015	1677	79
SRX743913	MET0241	HV11	Female	20s	Hp-R	22528504	31678	22929	2265	25194	22471632	99.75	44943264	106734	124822716	200	440822	1169	2175	100
SRX744006	MET0259	HV12	Male	20s	Hp-R	96800668	163669	45190	5509	50699	96586300	99.78	193172600	313045	268674794	200	460331	858	1114	187
SRX744575	MET0276	HV13	Male	20s	Hp-R	86617215	148445	40251	5484	45735	86423035	99.78	172846070	246676	198367753	200	483531	804	1009	172
SRX744593	MET0293	HV14	Male	30s	Hp-R	48404628	44950	33602	3244	36846	48322832	99.83	96645664	138201	153663452	200	679438	1112	2501	75
SRX745020	MET0310	HV15	Female	30s	Hp-R	48472837	93642	34345	3158	37503	48304350	99.65	96608700	298410	294144005	200	551747	986	1454	110

Table 3 Repeat sequences commonly detected in skin metagenome datasets.

Sequence	Hit by BLAST search	Repeat ID	Max score	Total score	Query cover	E value	Ident (%)	Frequency	# of Samples Observed In	Potential Cas Type	Potential direction
GTTTATAGAGCTGTGTTGTTTCGAATGGTCCAAAAC	<i>Streptococcus equinus</i> strain NCTC11436 genome assembly, chromosome: 1	STEQ* ¹	66.2	1945	100%	1E-08	100.00	157	8	IIA	-
CCAGCCGCCCTCAGGCGGCTGTGTGTTGAAAC	<i>Neisseria polysaccharea</i> strain M18661 chromosome, complete genome	NEPO	59	4750	100%	0.000001	100.00	273	6	IC	ND
GTTTTGTACTCTCAAGATTTAAGTAAGTGTACAAC	<i>Streptococcus thermophilus</i> isolate NWC_1_1 chromosome, complete genome	STTH* ²	66.2	1982	100%	1E-08	100.00	78	5	IIC	-
GTTTCAATACCGTAAACGGTAAGCACTGTTGCAAC	<i>Veillonella parvula</i> strain NCTC11810 genome assembly, chromosome: 1	VEPA	64.4	2183	100%	3E-08	100.00	97	5	IIID	-
GTTGTGAATTGCTTTCAAAAATTATTATCTTCGTGATGAAGGACAAC	<i>Chryseobacterium</i> sp. H6466 chromosome, complete genome	CHXX	86	1199	100%	4.00E-14	100.00	105	4	IIC	ND
GCTGGGGTTACAGTTCTCATTACCCCTGATAGACTTC	<i>Corynebacterium diphtheriae</i> strain NCTC7838 genome assembly, chromosome: 1	CODI	52.7	368	100%	0.0003	91.67	68	4	IIU	ND
GTCTCCGCTCTGTGGGAGCGGACTTCGTTGAGGC	<i>Micrococcus luteus</i> strain SGAir0127 chromosome, complete genome	MILU	66.2	1950	100%	1.00E-08	100.00	91	3	IU	ND
GTTGTAGCTCCCTTTTTCATTTCGCAGTGCTATAAT	<i>Haemophilus parainfluenzae</i> strain LC_1315_18 chromosome, complete genome	HAPA	66.2	182	100%	1E-08	100.00	73	3	IIU	ND
GTCAGAAAGCACCCAGCACAGAGGTGCATTAAGAC	<i>Actinomyces oris</i> strain T14V, complete genome	ACOR	68	2980	100%	3E-09	100.00	33	3	ND	ND
GTTTTAGAGGTTTACAGAATTACACTACTCTCAAAC	<i>Veillonella rodentium</i> strain NCTC12018 genome assembly, chromosome: 1	VERO	61.7	1689	100%	5E-07	97.22	19	3	IIA	ND
GCAGCCACCTTCGGGTGGCTGTGTGTTGAAAC	<i>Aggregatibacter aphrophilus</i> strain NCTC11096 genome assembly, chromosome: 1	AGAP	59	530	100%	0.000001	100.00	20	3	IC	ND
GTCGCACCCACACGGGTGCGTGGATTGAAAC	<i>Syntrophomonas wolfei</i> subsp. <i>wolfei</i> str. Goettingen G311, complete genome	SYWO	54.5	3098	100%	0.00004	96.88	164	3	ND	+
GTTTCAATCCACCACCTAGTCGAAACCAGGTGGAAAC	NA	NA	NA	NA	NA	NA	NA	42	2		
GTTTCAATTAACAGCCTCTCCGTAGAAAGGCTGCAACC	<i>Alistipes</i> sp. Marseille-P5997 genome assembly, chromosome: contig00001		37.4	37.4	71%	6.9	92.59	8	2		
GTCGCAGCCTCTCTCCGAGGGGCTGAGGATTGAAAC	<i>Mesorhizobium</i> sp. M6A.T.Cr.TU.016.01.1.1 chromosome		41.9	41.9	67%	0.47	96.00	16	2		
GTTTCAATCCACGCTCCCGCTAAGGAGCGAC	Uncultured bacterium gwf2_scaffold_15186 CRISPR-Cas system sequence		54.5	871	100%	0.00004	96.88	14	2		
GTTTCCGTCCCCTCTCGAGGTAGATAGGTTCTTATA	<i>Streptococcus oralis</i> strain S.MIT/ORALIS-351, complete genome		66.2	992	100%	1E-08	100.00	10	2		
GTAAGAAACCCAGTCACCTCGGGAGGGGACGAAAC	<i>Streptococcus sanguinis</i> SK36, complete genome		68	1555	100%	3E-09	100.00	9	2		
GTTTCCGTCCCCTCTCGGGGATTAAGGCATCTGAAC	<i>Mycobacterium tuberculosis</i> RGTB327, complete genome		44.6	1424	94%	0.039	88.24	12	2		
GTTGGGAATCTTCACTAACTATCCTTGCTAGACTTA	<i>Rothia dentocariosa</i> ATCC 17931, complete genome		42.8	42.8	77%	0.13	92.86	9	2		
GTTTTCGTCCCCTCTCGGGAGTAAGGTATCTGAAC	<i>Mogibacterium pumilum</i> strain ATCC 700696 genome		37.4	373	55%	5.7	100.00	6	2		
GTCGCACCCCTTCGCGGGTGCCTGGATTGAAAT	<i>Streptococcus</i> sp. Z15 chromosome, complete genome		59	59	100%	0.000001	100.00	9	2		
GTTTCAACACGCAGCCGCCCGAAGGCGGCTGG	<i>Neisseria elongata</i> strain M15911 chromosome, complete genome		59	532	100%	0.000001	100.00	11	2		
GGTTTTGTAGTCTCCAAAATTGTGACCGATAAAAC	<i>Streptococcus pneumoniae</i> strain 2245STDY6020221 genome assembly, chromosome: 1		66.2	66.2	100%	1E-08	100.00	7	2		

*1: Referred to as “CRISPR 3” in chapter II.

*2: Referred to as “CRISPR 1” in chapter II.

Table 4 Summary of CRISPRCasFinder results.

CRISPR_ID	CRISPR_Start	CRISPR_End	CRISPR_Length	Potential_Orientation (A1%)	Consensus_Repeat	Repeat_ID	Spacers_Nb	Mean_size_Spacers	Nb_Repeats_matching_Consensus	Conservation_Repeats (% identity)	Conservation_Spacers (% identity)	Repeat_Length_plus_mean_size_Spacers	Evidence_Level	Cas
SRX740541-k99_14457_1	1701	0	581	1628	1047 Unknown	TTTCTTAGCTGCCTATACGGCAGTGAAC	17	32.0	15	71.4	0.0	60.0	4	CAS-TypelF[17;457],
SRX740541-k99_34147_1	2064	0	1427	2004	577 Unknown	GTGTTCCCCACATCGCTGGGGATGAACCG	9	32.0	10	100.0	0.0	61.0	4	CAS-TypelE[117;1362],
SRX740541-k99_4199_1	6443	0	5535	6403	868 Unknown	GTTTCATGGCGGCATACGCCATTAGAAA	14	32.1	15	100.0	0.0	60.1	4	CAS-TypelF[1320;5402],
SRX740589-k99_19418_1	2447	0	1	501	500 Unknown	GTTTCGTCGCCCATCAGGGGATTATTTAATTTAAC	6	42.7	7	100.0	0.0	77.7	4	CAS[819;2091],
SRX740632-k99_26903_1	2756	0	2070	2699	629 Unknown	TTTTAGAGCTGTGTTGTTTTCGAATGTTCCAAAAC	9	30.0	10	100.0	0.0	66.0	4	CAS[120;1979], CAS-TypelA[1317;1979], CAS-TypelIU[120;1979],
SRX740632-k99_60691_1	1608	0	909	1606	697 Unknown	GCAGCCGCCTTCGGGCGGCTGTGTTGTGAAC	10	34.6	11	100.0	0.0	66.6	4	CAS[424;717],
SRX741396-k99_95278_1	5623	0	4925	5564	639 Unknown	GTGCTCCCGCGGTGAGCGGGGATGAGCC	10	33.2	11	100.0	2.0	61.2	4	CAS-TypelE[277;4839],
SRX743869-k99_125839_1	3642	0	1455	3069	1614 Reverse	GGAAATACCTCCGCGTGC GCGGAGAGAG	26	32.0	24	86.2	0.0	61.0	4	CAS-TypelE[3088;3402],
SRX743869-k99_20198_1	11050	0	6316	7077	761 Reverse	GTTTTAGAGCTGTGCTGTTTTCGAATGTTCCAAAAC	11	30.0	12	100.0	0.0	66.0	4	CAS[221;6215], CAS-TypelA[5550;6215], CAS-TypelIU[221;6215],
SRX743869-k99_45857_1	15116	0	108	623	515 Reverse	GGCTCATCCCCGCTACGCGGGGAGCAC	8	33.0	9	100.0	0.0	61.0	4	CAS[710;9885], CAS-TypelE[710;7008],
SRX743869-k99_7556_1	10355	0	9774	10350	576 Unknown	GTGCTCCCGCGCTAAGCGGGGATGAGCC	9	33.0	10	100.0	0.0	61.0	4	CAS[509;9686], CAS-TypelE[3386;9686],
SRX743869-k99_95278_1	4873	0	17	656	639 Unknown	AGGGCTCATCCCCGATGCGCGGGGAAAC	10	31.0	4	93.3	0.0	61.0	4	CAS-TypelE[737;4869],
SRX743869-k99_9627_1	13578	0	10288	10983	695 Reverse	GTTTTAGAGCTGTGTTGTTTTCGAATGTTCCAAAAC	10	30.0	10	94.4	4.5	66.0	4	CAS[4160;10197], CAS-TypelA[9535;10197], CAS-TypelIU[4160;10197],
SRX743913-k99_44863_1	19595	0	13692	17684	3992 Forward	GTTTCGCTCGCGCACAGGCAGCTCAGAAA	66	32.1	62	57.1	0.0	60.1	4	CAS-TypelF[5159;13561],
SRX743913-k99_91515_1	7731	0	6404	7690	1286 Unknown	TTTTTTGTACTCTCAAGATTTAAGTAAGTGTACAAC	19	29.8	20	100.0	0.0	65.8	4	CAS[474;5293], CAS-TypelIU[474;4965],
SRX744006-k99_132483_1	1376	0	52	348	296 Unknown	GTTTCAACACACAGCAGCCGGAAGGCGGCTGG	4	34.3	4	96.9	6.8	66.3	4	CAS[538;831],
SRX744006-k99_146201_1	4760	0	4214	4760	546 Unknown	CGTGTCTCCGGTGTAAAGCCGAGGCTCCATTGCGGG	7	35.9	6	97.3	2.2	72.9	4	CAS[68;4044], CAS-TypelD[2666;3757],
SRX744006-k99_17893_1	6883	0	6366	6863	497 Unknown	GTTTTAGAGCTGTGTTGTTTTCGAATGTTCCAAAAC	7	30.0	8	100.0	2.2	66.0	4	CAS[280;6275], CAS-TypelA[5613;6275], CAS-TypelIU[280;6275],
SRX744006-k99_185168_1	4901	0	515	945	430 Unknown	CCAGCCGCCTTCGGGCGGCTGTGTTGAAACA	6	33.3	4	72.7	2.1	66.3	4	CAS[32;325],
SRX744006-k99_193588_1	7413	0	18	501	483 Unknown	GTTGGATTCTTTCTAAATACTCCTTGATATACTTA	7	28.0	8	100.0	2.8	64.0	4	CAS[676;1897], CAS-TypelIU[986;1897],
SRX744006-k99_246214_1	2048	0	1546	2043	497 Unknown	ATTATAGCACTGCGAAATGAAAGGGAGCTACAAC	7	30.0	8	100.0	2.6	66.0	4	CAS[247;1483], CAS-TypelIU[247;1164],
SRX744006-k99_277696_1	12685	0	10079	10511	432 Reverse	ATTGTAGCACTGCGAAATGAGAAAGGGAGCTACAAC	6	30.2	6	81.1	2.9	66.2	4	CAS[5470;10015], CAS-TypelIU[5470;9696],
SRX744006-k99_300010_1	2584	0	1187	1581	394 Reverse	ATTTCCCGCCTTCGGGCGGGGACTTTTCATTGAGGC	5	35.8	5	81.1	0.0	71.8	4	CAS[595;894],
SRX744006-k99_309270_1	8422	0	7468	8411	943 Unknown	GTGCTCCCGCAGACGCGGGGATGAGCC	15	33.1	16	100.0	0.0	61.1	4	CAS-TypelE[1368;7423],
SRX744006-k99_31955_1	1436	0	837	1397	560 Unknown	GCAGCCGCCTCCGCGCGGCTGTGTTGAAAC	8	34.1	9	100.0	1.6	66.1	4	CAS[352;645],
SRX744006-k99_7713_1	4840	0	4045	4561	516 Reverse	GTCAGCCCGCGCATGCGGGGGTAGAGG	8	33.1	3	64.3	0.0	61.1	4	CAS-TypelIU[1506;2717],
SRX744006-k99_78272_1	8929	0	2854	3327	473 Reverse	CCCTCAATGAAAGTACCCATTTTCATGGGTGAGAC	6	37.0	4	88.9	5.7	73.0	4	CAS[3522;8905], CAS-TypelIU[7919;8905],
SRX744006-k99_81987_1	5386	0	1	367	366 Unknown	GTTTTAGAGGTTTACAGAAATACACTACTCTCAAAC	5	30.2	6	100.0	2.6	66.2	4	CAS[432;5214], CAS-TypelA[432;1106], CAS-TypelIU[432;5214],
SRX744575-k99_13938_1	1273	0	9	377	368 Unknown	GTTTCAACACACAGCCGCTAAGGCGGCTGG	5	35.4	6	100.0	4.2	67.4	4	CAS[567;860],
SRX744575-k99_163833_1	13840	0	12207	13778	1571 Unknown	GAAGTCTATCAGGGGTAATGAGAACTGAACCCAGC	24	28.0	25	100.0	0.0	64.0	4	CAS-TypelIU[10926;11846],
SRX744575-k99_1683_1	20838	0	2041	2434	393 Forward	GTTTTCCCGCCTACGCGGGGTGAGCC	6	33.0	7	100.0	0.0	61.0	4	CAS-TypelE[29;1986],
SRX744575-k99_3867_1	15289	0	12317	15272	2955 Unknown	GTGCTCCCGCGGTGAGCGGGGATGAGCC	48	33.0	47	92.9	0.0	61.0	4	CAS[3017;12232], CAS-TypelE[5928;12232],
SRX744575-k99_41114_1	12204	1	41	2023	1982 Unknown	GGGCTCATCCCGCGGGTGC GGGGCAAC	32	32.1	24	93.1	0.0	61.1	4	CAS[2283;11405], CAS-TypelE[5160;11405],
SRX744575-k99_58530_1	3712	0	1088	1480	392 Reverse	GTTTCAATCCACCACCTAGTCGAAACAGGTGGAAC	5	34.2	4	91.9	7.0	71.2	4	CAS[334;3567], CAS-TypelC[1948;2976],
SRX744575-k99_75278_1	14891	0	1633	1780	147 Forward	AACCCCGCATACGCGGGGAAAGAG	2	37.5	0	70.8	53.8	61.5	1	CAS[3706;10912], CAS-TypelE[6322;10912],
SRX744575-k99_75278_2	14891	0	3068	3211	143 Reverse	AACCCCGCATACGCGGGGAAAGAG	2	35.5	0	75.0	51.2	59.5	1	
SRX744575-k99_75278_3	14891	0	3310	3637	327 Reverse	AACCCCGCATACGCGGGGAAAGAG	5	36.8	1	66.7	10.6	60.8	3	
SRX744575-k99_75278_4	14891	0	10980	11673	693 Forward	CTTTTCCCGCGCGTGC GGGGTTGTTTC	11	32.5	3	78.6	0.0	60.5	4	
SRX744575-k99_8708_1	12432	0	51	1908	1857 Unknown	GGCTCATCCCGCTCAGCGGGGAGCAC	30	33.0	31	100.0	0.0	61.0	4	CAS[2213;11306], CAS-TypelE[5116;11306],
SRX744593-k99_124510_1	2516	0	42	434	392 Unknown	GGTCGACGCTCTCTCCGAGGGGCTGAGGATTGAAAC	5	33.0	3	94.7	5.0	71.0	4	CAS[546;2455], CAS-TypelC[1137;2165],
SRX744593-k99_27333_1	3244	0	580	3188	2608 Unknown	GTTTCATGGCGGCATACGCCATTAGAAA	43	32.0	44	100.0	0.0	60.0	4	CAS-TypelF[28;447],
SRX745020-k99_123677_1	2865	0	2345	2794	449 Unknown	GTTTACTGCGGTGCAGCGAGCTTAGAAAAT	7	30.0	3	83.3	0.0	60.0	4	CAS-TypelF[241;2076],
SRX745020-k99_185373_1	11716	0	3	930	927 Unknown	TTTCTAAACCATCTATGCGATGGCGAAC	15	32.0	16	100.0	0.0	60.0	4	CAS-TypelF[1028;9452],
SRX745020-k99_226394_1	6396	0	339	698	359 Reverse	GTTGTGAGTTCCCTTTCACTTTTGTAGTAATTTTACGTAAAGTAGTTACAAC	4	30.0	5	100.0	5.7	78.0	4	CAS[805;5596], CAS-TypelIU[1140;5596],
SRX745020-k99_32511_1	22370	0	8936	9818	882 Forward	GTGTTCCCGCGCTAGCGGGGATAAACCG	14	32.0	15	100.0	0.0	61.0	4	CAS[389;8839], CAS-TypelE[3061;8839],

Table 5 Numbers of reads containing the CRISPR repeat in the samples.

ID	Total initial reads pair	STEQ	NEPO	STTH	VEPA	CHXX	CODI	MILU	HAPA	ACOR	VERO	AGAP	SYWO
HV15_Hp-R	48472837	114	443	162	360	1759	10	89	66	73	4	449	1070
HV03_Hp-R	8650474	59	0	41	4	523	0	0	0	63	0	0	7
HV02_Hp-R	2353252	6	6	27	2	224	2	14	0	4	2	2	19
HV14_Hp-R	48404628	150	153	198	23	2961	134	42	27	41	2	149	234
HV13_Hp-R	86617215	338	967	159	203	15	1325	1200	52	66	131	272	1046
HV01_Hp-R	2195791	0	1	3	0	0	2	7	0	0	0	0	4
HV07_Hp-R	8777711	92	51	30	29	4	0	204	8	66	22	42	190
HV07_N	62115237	2	0	0	0	0	0	12	0	0	0	0	0
HV07_Hp-L	19702061	38	44	8	3	4	0	99	2	16	6	37	39
HV07_Ch	46861012	17	10	10	7	0	0	52	1	15	12	10	20
HV06_Hp-R	11836038	495	2409	420	54	0	2	59	209	213	134	1923	46
HV09_Hp-R	362229982	3028	4994	1231	315	38	370	2042	1128	1201	122	4880	1300
HV08_Hp-L	3375857	4	4	0	0	0	0	0	0	0	0	0	0
HV12_Hp-R	96800668	2570	2126	873	351	17	162	901	561	496	241	1617	493
HV10_Hp-R	28464770	487	220	179	96	57	102	20	337	31	5	206	897
HV05_Hp-R	5993912	215	6	120	105	0	0	33	19	114	72	1	1
HV11_Hp-R	22528504	729	212	575	141	318	0	13	31	56	125	199	367

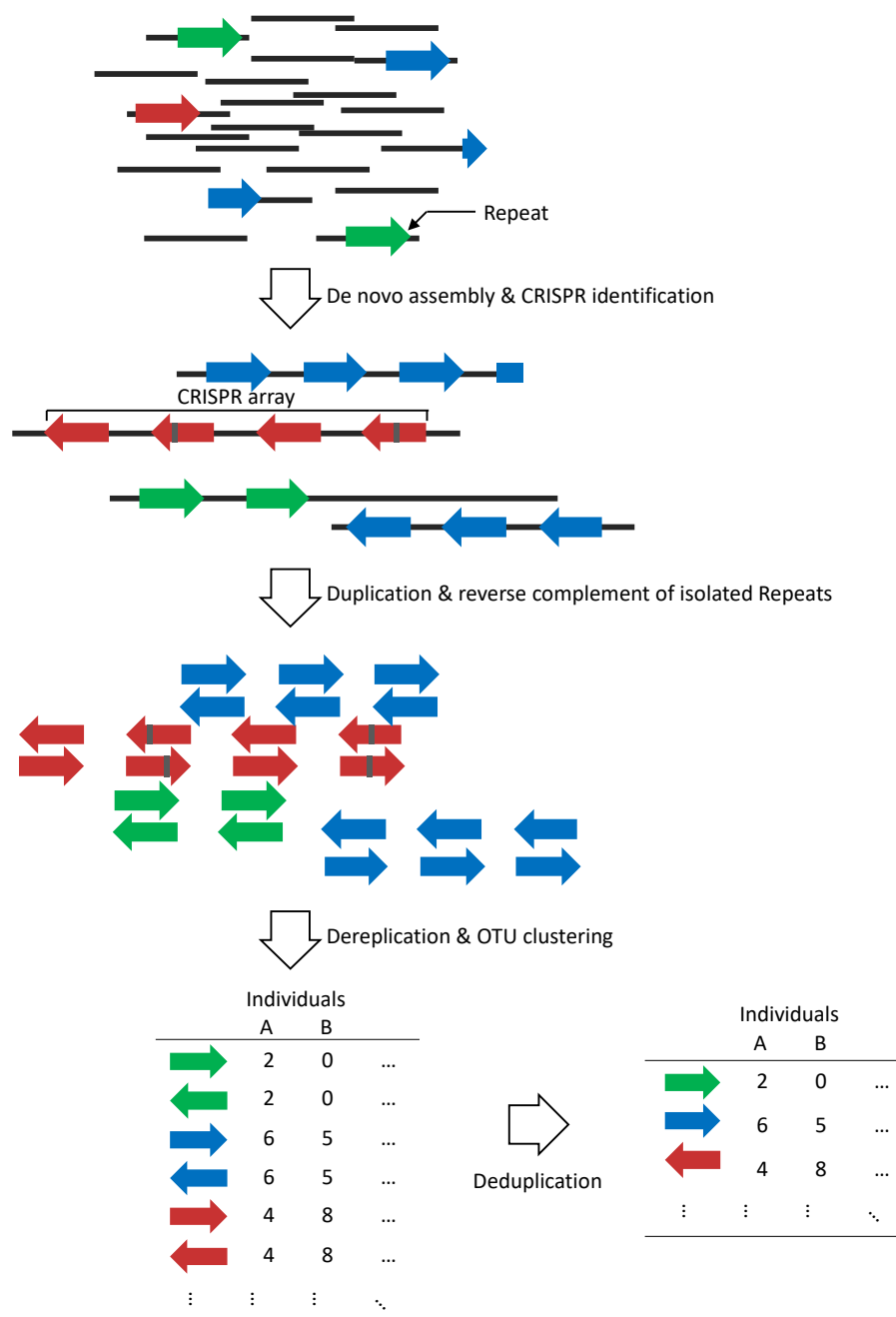


Figure 1 Schematic of CRISPR reconstruction from the metagenomic reads and identification of shared CRISPR repeats.

De novo assembly of the metagenome was performed with MEGAHIT v1.1.4. Output contigs were analyzed using PILER-CR v1.06 with default parameters to detect the CRISPR arrays. All reads were then duplicated, and one of the read pairs was reverse-complemented. Repeat reads were then dereplicated and clustered into OTUs by QIIME 2 v2019.7. The duplicated reverse complement repeat reads were discarded after OTU clustering.

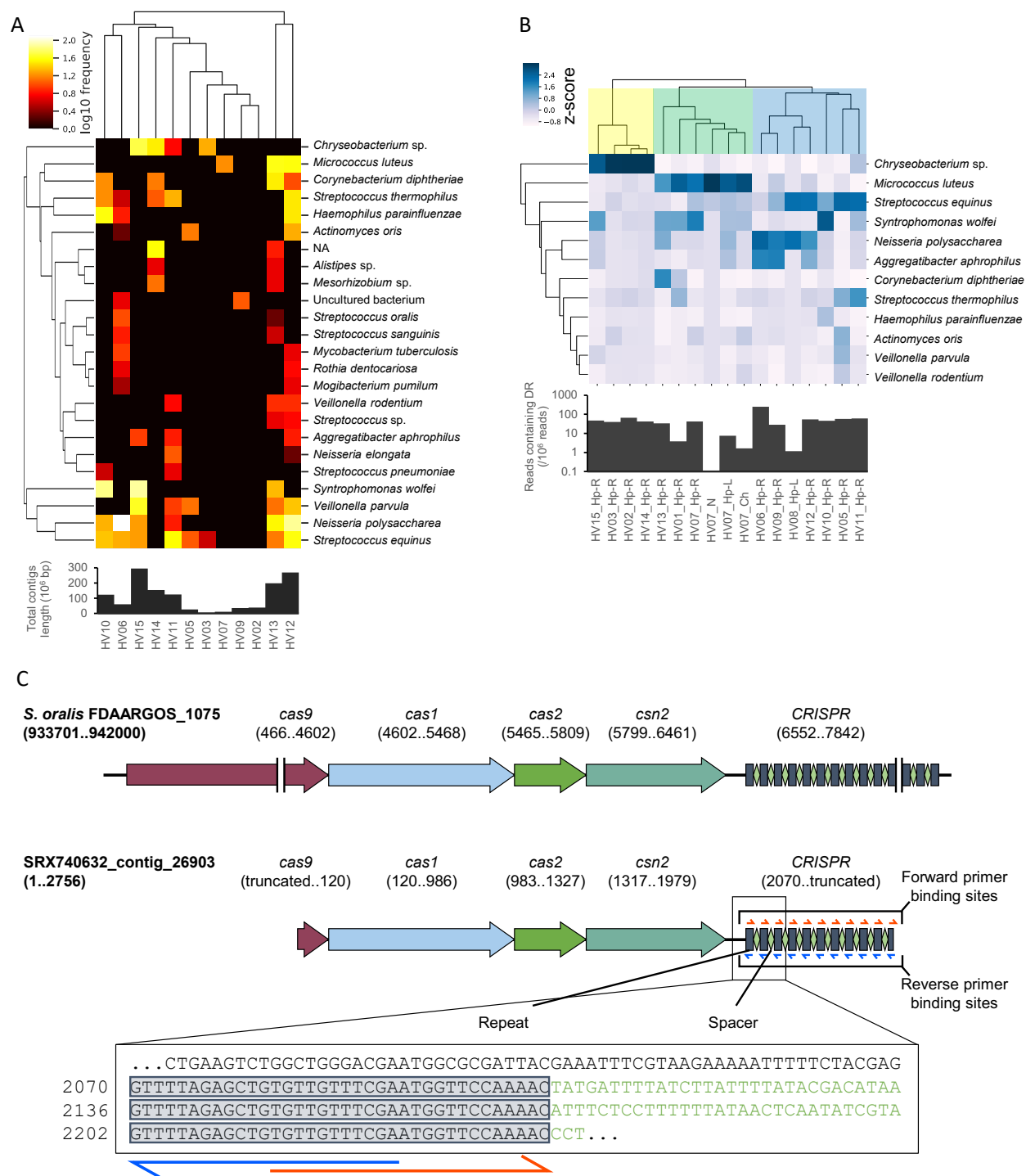


Figure 2 Metagenomic reconstruction of CRISPR arrays identified commonly shared repeats in skin metagenome datasets.

(Continued)

(A) Heatmap of the repeat matrix depicting the abundance of each repeat derived from putative CRISPR arrays detected in metagenomic data of the bacterial community in human skin. The bar chart under the heatmap shows the total contig length after de novo assembly. (B) Heatmap indicating the abundance of CRISPR repeats in metagenomic datasets detected by BLAST search against SRA data. Each column was standardized by the z-score. The bar chart under the heatmap shows reads containing repeats used as queries in the BLAST search. Hp, palm; N, nares; Ch, cheek. (C) Representative putative STEQ CRISPR-Cas system detected by metagenomic reconstruction. CRISPR-Cas locus of *Streptococcus oralis* strain FDAARGOS_367 (GenBank: CP023507) is shown as the reference. The presence of the characteristic *csn2* gene in the locus suggests that this system is type II-A CRISPR-Cas system; the *cas9* gene, the effector protein of type II system, was partially detected in the reconstructed CRISPR array because 5'-end of the entire CRISPR-Cas locus was truncated. The box below indicates the partial sequence of the CRISPR array. The marked sequences are repeat sequences, while the sequences colored in green are spacer sequences. Forward and reverse primer binding sites on the repeats are indicated by red and blue arrows, respectively.



Figure 3 Alignment of the representative metagenomically reconstructed Streptococcal CRISPR arrays against reference CRISPRs.

Spacers of three STEQ CRISPRs detected by CRISPRCasFinder and 26 Streptococcal reference CRISPRs were visualized by CRISPR Visualizer (60, 62). Each colored box represents a spacer. Grey boxes with “x” represent missing spacers. Note that similar colors do not indicate similar sequences. All metagenomically reconstructed spacers are unique among the 26 arrays.

CHAPTER II

Evaluation of CRISPR diversity in the human skin microbiome for personal identification

Introduction

CRISPR is a genetic region of prokaryotes, which is composed of conserved repeats and hypervariable spacers. As the variation of spacer sequences reflect history of phage infection and are distinct for each strain, CRISPRs provide phylogenetic information that is useful for strain-level typing of pathogenic bacteria, such as *Mycobacterium* spp. (33). Such characteristics are also used to visualize host-phage interaction by amplicon sequencing of CRISPR loci (32, 44, 48-50), suggesting that PCR-based CRISPR analysis could be utilized for personal identification which requires high resolution classification.

In chapter I, 24 putative CRISPR loci were identified by metagenomic reconstruction of CRISPRs in human skin microbiome. However, it would be difficult to capture spacer diversity comprehensively by an assembly-based approach. Moreover, metagenome analysis is difficult to use routinely for forensically relevant samples due to its limitations, including analytical complexity and high cost. Thus, it is necessary to develop a new PCR-based typing method.

The studies on CRISPR diversity utilizing amplicon sequencing are limited until now. Previous reports analyzed *Streptococcus* CRISPRs in order to analyze host-phage interaction in oral microbiome (32, 48-50, 68, 69). Interestingly, metagenomic

reconstruction performed in chapter I revealed that *Streptococcus* CRISPRs are the most abundant CRISPRs in human skin environments. As *Streptococcus* is a most abundant genus of skin microbiome and streptococcal CRISPRs are also analyzed in the previous study (50), streptococcal CRISPRs could be good targets for PCR-based personal microbiome typing rather than conventional 16S rRNA sequencing, the resolution of which is limited to genus-level.

In this chapter, the spacers of three putative CRISPR arrays in the skin microbiome were sequenced to characterize the diversity of spacer sequences. In addition to two streptococcal CRISPRs, a CRISPR locus from *Micrococcus luteus*, which is also shown to be another abundant CRISPR in chapter I, was analyzed for comparison. The diversity of spacer sequences was evaluated by calculating the alpha diversity index and the beta diversity index. Then, it is evaluated whether CRISPR array sequencing can be used as a novel tool in personal identification by using machine learning approach. This CRISPR based analysis was validated by comparing the pipeline with 16S rRNA sequencing.

Materials and methods

Sample collection

Five individuals were asked to participate in this study and swab samples were collected for spacer sequencing. The participants provided information regarding their gender, age range, systemic antibiotics usage in the past six months, steroid usage in the past six months, skincare product usage in the past 24 h, and handwashing in the past 2 h. Table 6 summarizes the detailed information of samples. Skin surfaces (approximately 5

× 5 cm) were swabbed with cotton-tipped swabs soaked in water. The hypothenar palm (Hp), antecubital fossa (Ac), and retroauricular crease (Ra), representing dry, moist, and sebaceous sites, respectively, were selected for swabbing (12). For the initial sampling (timepoint 1), only Hp sample was obtained, except for one individual (P05). Two to three months later (timepoint 2), three skin sites (Hp, Ac, and Ra) were sampled for all individuals. After swabbing, the cotton tips were separated and placed in 2 ml tubes. A pre-moistened swab tip placed directly in the 2 ml tube was used for a blank for DNA extraction (EC). Samples were stored at -20 °C before DNA extraction. All procedures involving human participants were approved by the Institutional Ethics Committee of the National Research Institute of Police Science.

DNA extraction

Genomic DNA was extracted from swabs using the DNeasy PowerSoil Kit (Qiagen, Aarhus, Denmark) with some minor modifications (11, 70). Briefly, frozen cotton tips were transferred into the bead tube provided with the DNeasy PowerSoil Kit, followed by incubation at 70 °C for 15 min after addition of the C1 solution. Tubes were then horizontally shaken at 3200 rpm for 2 min on a bead beater-type homogenizer μ T-12 (TAITEC, Saitama, Japan). Remaining steps were performed following the manufacturer's instructions. Genomic DNA was eluted in 100 μ l of the C6 solution.

Amplification and sequencing of spacers

To amplify the spacer regions of the putative CRISPRs, three pairs of primers were designed based on the specificity of the CRISPR repeat, namely STEQ, STTH, and MILU (Table 7). These primer pairs are specific to the 3' region of each CRISPR repeat, so that the primer pair flanks a spacer. The length of primers was determined to meet the

melting temperature requirement on the Illumina 16S Metagenomic Sequencing Library Preparation protocol (https://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf). PCR was performed in a 50- μ l reaction volume using the KOD FX (Toyobo, Osaka, Japan). Each reaction consisted of 1 $\times$ PCR Buffer for KOD, 0.4 mM dNTPs, 0.3 μ M of each primer, 1.0 U of KOD FX, and 2 μ l of the DNA template. EC and a non-template control (NTC) for PCR were included as negative controls. The following conditions were used for amplification: 2 min at 94 °C, 35 cycles of denaturation at 95 °C for 30 s, annealing at 60 °C for 30 s, and extension at 72 °C for 30 s, followed by final extension at 72 °C for 5 min. Next, 25 μ l of the products was purified using 1.8 $\times$ volume of the Agencourt AMPure XP kit (Beckman Coulter, Brea, CA) twice and washed using freshly prepared 80% ethanol. DNA was eluted in 52.5 μ l of 10 mM Tris (pH 8.5) and 0.1% Tween 20 elution buffer. If low-molecular-weight bands were observed using agarose gel electrophoresis, additional purification was performed using 1.8 $\times$ volume of the Agencourt AMPure XP kit. Purified DNA was indexed using the Nextera XT DNA index kit (Illumina, San Diego, CA,) and KOD FX in a reaction volume of 50 μ l. Each reaction consisted of a 1 $\times$ PCR Buffer for KOD, 0.4 mM dNTPs, 4.5 μ l of each index primer, 1.0 U of KOD FX, and 5 μ l of the purified DNA as template. The index PCR product was purified using 1.8 $\times$ volume of the Agencourt AMPure XP kit and eluted in 27.5 μ l of 10 mM Tris (pH 8.5) and 0.1% Tween 20 elution buffer. The indexed libraries were quantified using GenNext NGS Library Quantification Kit (Toyobo) and then diluted to 4 nM with an elution buffer. If the library concentration was less than 4 nM, the library was pooled with other libraries without dilution. Denaturing and final dilution were performed according to the manufacturer's instructions. Each run

included > 15% of the PhiX control v3 (Illumina) to improve sequencing quality. Paired-end sequencing of 300 cycles (2×151 bp) was performed on a MiSeq platform (Illumina) with the MiSeq v2 chemistry, according to manufacturer's instructions.

Amplification and sequencing of 16S rRNA

Sequence libraries of the 16S rRNA were prepared according to the Illumina 16S Metagenomic Sequencing Library Preparation protocol with some modifications. The primer pair targeting V1-V3 region (71, 72) was used in this study (Table 7). PCR was performed in a 25- μ l reaction volume using the KAPA HiFi HotStart ReadyMix kit (KAPA Biosystems, MA). Each reaction consisted of 1 $\times$ KAPA HiFi HotStart ReadyMix, 0.3 μ M of each primer, and 2 μ l of the DNA template. The following conditions were used for amplification: 5 min at 95 $^{\circ}$ C, 25 cycles of denaturation at 95 $^{\circ}$ C for 30 s, annealing at 62.3 $^{\circ}$ C for 30 s, and extension at 72 $^{\circ}$ C for 30 s, followed by final extension at 72 $^{\circ}$ C for 5 min. Then, 25 μ l of the products was purified using 0.8 $\times$ volume of the Agencourt AMPure XP kit and washed using freshly prepared 80% ethanol. DNA was eluted in 52.5 μ l of 10 mM Tris-HCl (pH 8.5) and 0.1% Tween 20 elution buffer. Purified DNA was indexed using the Nextera XT DNA index kit (Illumina) and the KAPA HiFi HotStart ReadyMix kit in a reaction volume of 50 μ l. Each reaction consisted of a 1 $\times$ KAPA HiFi HotStart ReadyMix, 5 μ l of each index primer, and 5 μ l of the purified DNA as template. The index PCR product was purified using 1.1 $\times$ volume of the Agencourt AMPure XP kit and eluted in 27.5 μ l of 10 mM Tris (pH 8.5) and 0.1% Tween 20 elution buffer. The indexed libraries were quantified by the GenNext NGS Library Quantification Kit and diluted to 4 nM with an elution buffer. If the library concentration was less than 4 nM, the library was pooled with other libraries without dilution. Denaturing and final

dilution were performed according to the manufacturer's instructions. Each run included > 15% of the PhiX control v3 to improve sequencing quality. Paired-end sequencing of 600 cycles (2×301 bp) was performed on a MiSeq platform (Illumina) with the MiSeq v3 chemistry, according to the manufacturer's instructions.

Data processing for spacer analysis

Initial quality controls were carried out using the CLC Genomics Workbench 12.0.2 (Qiagen). After low-quality sequences ($Q < 20$) and primer sequences were removed, reads shorter than 20 bp or longer than 40 bp were filtered. The remaining spacer sequences were then analyzed using QIIME 2 v2019.7 (54). Spacer reads were denoised and dereplicated with the denoise-paired method of q2-dada2 plug-in (73) with `min_fold_parent_over_abundance = 32`; the resulting amplicon sequence variant (ASV) tables were further analyzed. Rarefaction curves were created with the q2-diversity plug-in with 100 iterations and 40 steps at 10000 reads of maximum depth. Based on rarefaction curve analysis, each sample was rarefied to 4829 reads for normalization in further analysis and samples with less than 4829 spacer ASVs were omitted from the calculation (74). The EC and NTC were removed at this step; however, EC was included in STEQ spacer analysis. Shannon indices were calculated in bits with the q2-diversity plug-in and Kruskal–Wallis tests were performed to compare Shannon indices between samples (75). For calculation of the beta-diversity index, the Bray-Curtis dissimilarity index was calculated by the q2-diversity plug-in and used for Principal coordinates analysis (PCoA) plots. Mann–Whitney U test was calculated using the “stats.mannwhitneyu” function of the python library “SciPy” (version 1.3.0) to compare the Bray-Curtis dissimilarity indices. Hill numbers were calculated using the library

“vegan” (version 2.5-4) in R (version 3.4.4) (76).

To compare the CRISPR arrays in the samples, we reconstructed the arrays from spacer amplicons. The trimming and assembling described below were performed using the CLC Genomics Workbench 12.0.2. First, reads that were longer than 100 bp were extracted after low-quality sequences ($Q < 20$) were removed, so that reads contain two spacers per read. Then, these reads were de novo assembled by “Map reads back to contigs” mode with length fraction = 0.8 and similarity fraction = 0.9. These assembled contigs were then compared and visualized using CRISPR Visualizer (62).

The spacer sequences were subjected to BLAST analysis with default settings. If the reads had e-values of less than 0.01, they were further classified. If the description includes “phage” or “virus”, the query was classified as “Virus”. If the description includes name of host genus (“streptococcus” for STEQ or STTH and “micrococcus” for MILU), the query was classified as “Host”. Remaining queries with significant hits were classified as “Other”.

To identify a subset of predictive spacers and to improve interpretability of the classification, we performed feature selection using Support Vector Machine based on Recursive Feature Elimination (SVM-RFE). SVM-RFE is a machine-learning technique used to select a subset of features (e.g., genes associated with a specific disease), which are relevant to sample types from a large set of features, by recursively ranking the features and eliminating irrelevant features (77). In this study, SVM-RFE was implemented using the python library “Scikit-learn” (version 0.22.1) (78). For SVMs, the linear kernel was used with default parameters. Using SVM-RFE, 20 ASVs were selected from the dataset including all samples from an individual. Then, using the selected 20 features, the individuals were predicted by linear-kernel SVM using leave-one-out cross-

validation to measure predictive accuracy. However, nested cross-validation were not performed to optimize the selection of ASVs used in the linear-kernel SVM. This is because it is difficult to split the data twice for inner and outer cross validation due to the small sample size, which may cause overfitting of the data.

Data processing for 16S rRNA analysis

Reads were analyzed using QIIME 2 v2019.7 (54). Initially, 16S rRNA reads were denoised and dereplicated with the denoise-paired method of q2-dada2 plug-in (73). The parameters of trim_left_f = 20 and trim_left_r = 20 were set to trim primer sequences. The resulting ASV tables and representative reads were further analyzed. Representing reads were annotated by classify_sklearn methods of q2-feature-classifier plug-in with the SILVA 16S rRNA database v132 (78, 79). In this study, contaminants derived from Sphingomonadaceae and Pseudomonadaceae, which are known as common contaminants (80), were observed in both swab samples and NTC due to low biomass. Therefore, reads annotated as Sphingomonadaceae or Pseudomonadaceae were removed from the tables and representative reads using filter-table and filter-seqs methods of q2-taxa plug-in. The genus-level composition of the top 11 abundant ASVs were visualized as a bar plot using q2-taxa plug-in. For the calculation of the beta-diversity index, the Bray-Curtis dissimilarity index was calculated using the q2-diversity plug-in and then used for PCoA plots. Mann–Whitney U test was calculated using the “stats.mannwhitneyu” function of the python library “SciPy” (version 1.3.0). Feature selection and individual prediction were performed using SVM-RFE with linear kernel, as described above.

Data availability

Spacer sequences are available in the SRA database under accession numbers

DRA009650 and DRA010353. Meta 16S rRNA sequences are available in the SRA database under accession number DRA010342.

Results

Stability of evenness and richness of CRISPR spacer sequences in the skin microbiome

In chapter I, reconstruction of CRISPR arrays using metagenomic data detected putative CRISPR repeats commonly found in the skin microbiome. To test whether repeats detected in the reconstructed CRISPR arrays are also detected via PCR-based methods, three pairs of primers were designed in such a way that the primers flank the spacer and target the repeats of STEQ, STTH, and MILU, as they were frequently found during metagenomic reconstruction and are members of a major genus in the human skin (Figure 2C). CRISPR spacers were amplified with these primer pairs and sequenced using MiSeq. Primers targeting the repeats were used rather than primers flanking the entire CRISPR locus to amplify a broader range of CRISPR arrays (48). Note that the spacer amplicon is a mixture of spacers derived from multiple species since the same repeats are shared by several species. Thus, this approach cannot precisely determine which species are the origin of each spacer. Cutaneous skin swabs were collected from five individuals in their 20s to 40s and from the skin sites Hp, Ac and Ra, representing dry, moist, and sebaceous sites, respectively. The swab collections were conducted twice at a 2–3-month interval. DNA extracted from these swabs were used as a template for library construction; amplicon sequencing was performed using an Illumina MiSeq sequencer. The spacer reads were then trimmed to remove primer sequences; sequencing statistics

before and after primer trimming are summarized in Table 8. The average length of the spacer reads after primer trimming were 29.9, 29.7 and 35.3 for STEQ, STTH and MILU, respectively, which were comparable to typical spacer length and the spacer length observed in CRISPRCasFinder analysis (Table 4). These trimmed reads were then used as input for downstream analysis. After quality control implemented in QIIME2 (54), we detected a total of 476185 spacers and 664 unique spacers for STEQ, 460309 spacers and 759 unique spacers for STTH, and 286968 spacers and 204 unique spacers for MILU from 25 samples.

A rarefaction curve based on the observed ASVs reached a saturation plateau at a sequencing depth of 4000, indicating that the spacer diversities of all samples were well-represented at a relatively low sequencing depth (Figure 4A–C). Saturation of the rarefaction curve suggested low PCR error in the MiSeq results; therefore, an ASV table was used to evaluate spacer diversity rather than OTU clustering (81). Samples were rarefied to 4829 reads for normalization and those with lower numbers of reads were discarded from downstream analyses. Richness and evenness of the spacer diversity in each sample were evaluated by calculating the Shannon diversity index. The medians of Shannon diversity indices, after rarefaction and removal of low-depth samples, were 5.27 (interquartile range [IQR]: 4.32–5.81), 6.22 (IQR: 5.29–6.50), and 4.39 (IQR: 3.23–5.27) for STEQ, STTH, and MILU, respectively. The diversity profile of each CRISPR array was also estimated using Hill numbers, which reflect richness and evenness if plotted as a function of the parameter q (82). The Hill numbers plot indicated that STTH had the highest spacer richness and evenness (Figure 5). There were no significant differences in Shannon indices of the STTH and MILU spacers when we grouped samples by gender, age, or site characteristics. In contrast, significant differences in Shannon indices of the

STEQ spacers were observed when samples were grouped by gender (Figure 4D-F). These results indicate that the evenness and richness of spacer sequences in the skin microbiome are not affected by the environment from which the bacteria were collected.

Then, it was determined whether each spacer was specific to each individual or sample type characterized the distribution of each spacer sequencers among the individuals. In accordance with a previous report (48), it was found that the STEQ or STTH spacer sequences were shared within individuals even for samples collected from different skin sites or at sampling timepoint (Figure 6A and B), suggesting that these two CRISPRs are suitable for personal identification purposes. In contrast, the most abundant MILU spacers were shared between individuals, regardless of sample type (Figure 6C). Although microbial communities harboring the CRISPR arrays sequenced in this study were obtained from the same environment, a high degree of inter-individual CRISPR spacer dissimilarities were observed in STEQ and STTH but not in MILU. These results suggest that the frequency of formational change in CRISPR arrays, including spacer acquisition or loss, differs among CRISPRs even in the same environment.

CRISPR reconstruction from spacer amplicons

While the reads analyzed were limited to those containing a single spacer in a read, longer reads that contain two spacers in a read were also obtained. Using these reads, the CRISPR loci were also assembled from amplicons, as was done by Robles-Sikisaka et al. (50). Among the 136 assembled arrays, 35 arrays were found to share at least one spacer with another array; groups of arrays sharing spacers are shown in Figure 7. Among them, the STEQ CRISPR group 1 contains two identical CRISPRs from the same individual at different timepoints, suggesting a conserved CRISPR array on the skin.

Furthermore, it was observed that the CRISPR arrays shared identical spacers from the same individual at different sites on the skin at the same time point (e.g. STEQ group 2, STTH group 6, and MILU group 1), from the same individual at the same site on the skin at different time points (e.g. STEQ group 3, STTH group 4, and MILU group 2), or even from different individuals (e.g. STEQ group 1, STTH group 1 and MILU group 1), suggesting both conservation and diversity of these CRISPRs. However, it could not be determined whether these partial matches are derived from the acquisition/loss of spacers or assembling errors, especially when the spacers of the array ends did not match. Nevertheless, it was observed that the internal spacers were lost in some groups (STEQ group 1 and 6 and STTH group 7), suggesting that amplicon sequencing can be used to trace the phylogenetical relationship of CRISPR arrays in the environment.

Assessment of skin bacterial community by 16S rRNA sequencing

The structure of the skin microbiome was also assessed by using 16S rRNA V1-3 amplicon sequencing with a MiSeq sequencer. As a result, 1530894 reads were obtained from 25 samples after reads filtering (mean: 61235.76 reads/sample, IQR: 33802-90173). A total of 2,358 unique reads with a median length of 476 bp (IQR: 432-487) after primer trimming were obtained. Rarefaction analysis indicated that the number of observed species did not plateau at 5,000 reads, suggesting that 16S rRNA analysis requires more sequencing depth than spacer sequencing (Figure 4A–C, Figure 8). This relatively high number of required reads for 16 rRNA sequencing is consistent with a previous report that analyzed the V4-5 region (83). As shown in Figure 6D, the most common genus in the skin microbial community was *Cutibacterium*, which is represented by *C. acnes*, the most abundant species in the human skin (12, 13). Although *Moraxella* was specifically

and stably observed in individual P02, other abundant genera were observed in all individuals (e.g., *Cutibacterium*, *Staphylococcus*, and *Corynebacterium*) or specifically but not stably observed in an individual (e.g., *Acinetobacter* in individual P03). Thus, compared with CRISPR spacer diversity, the composition of abundant genera in the skin microbiome is likely insufficient for personal discrimination.

Putative protospacers of the spacers reads

Spacers of a CRISPR locus are acquired from previously encountered foreign genetic elements by host bacteria (28). To identify possible protospacers, which are the origins of spacers, spacer reads were subjected to BLAST searches (Figure 6E). For streptococcal CRISPRs (STEQ and STTH), 27.7% of spacer sequences were previously found in viruses, the sources of new spacers. In contrast, only one virus sequence was identified from MILU spacers. Some sequences were those of host bacteria; 15.1% for streptococcal spacers and 31.0% for MILU. These host-derived spacers may target plasmids, CRISPR loci, or are self-targeting (84). Interestingly, 34 sequences (16.7% of total unique spacers) of MILU spacers derived from the “host” were sequences of single strains of *M. luteus* (GenBank: CP025616.2). Spacers of the strain were regularly ordered with consistent repeats in the same direction, indicating that they were located on a single CRISPR locus. In total, 56.8% of the spacers were previously unknown sequences. This high percentage of unassigned reads was expected, as most phages have no homologous sequence in public databases (85). Altogether, the sequences identified by our spacer sequencing are likely derived from actual environmental CRISPRs, supporting the validity of our proposed method.

Individual-specific spacer patterns

The Bray-Curtis dissimilarity index for CRISPR spacers, which represents ecological similarities of the spacer community between samples, was calculated and compared with that of 16S rRNA V1-3 sequences; Bray-Curtis dissimilarity index = 1 if the two communities do not share any spacers, while Bray-Curtis dissimilarity index = 0 if the two communities are identical. PCoA of Bray-Curtis dissimilarity index of STEQ and STTH spacers showed strong clustering of the samples from the same individuals, indicating that the diversity of these spacers was more reflective of the host than of sample type (Figure 9A–F). In contrast, neither the PCoA plot of Bray-Curtis dissimilarity index of MILU spacers nor 16S rRNA showed any specific clustering (Figure 9G–L). Notably, STEQ and STTH spacers were significantly more individual-specific than site-specific (Figure 10A and B). No significant differences were observed in MILU spacers when the dissimilarities were compared among or between individuals (Figure 10C), indicating the differences in inter-individual dissimilarities of spacers among these CRISPRs. For STEQ and STTH, none of the Bray-Curtis dissimilarity indices calculated from a pair of inter-individual samples was <0.8 , indicating that these high inter-individual dissimilarities may sufficiently discriminate two individuals. This high dissimilarity in inter-individual samples was not observed for 16S rRNA sequencing (Figure 10D), suggesting that inter individual dissimilarity of microbiome measured by 16S rRNA sequencing does not have sufficient discriminating power for forensic cases. Collectively, these results indicate that STEQ and STTH spacer diversity in the skin microbiome is highly personalized compared to the microbiome diversity measured by 16S rRNA V1-3 sequencing; thus, this individuality can be used as a discriminating marker.

Personal identification by a set of predictive spacers

It was tested whether a set of discriminant spacers can aid identification of the personal microbiome by utilizing SVM-RFE (see Materials and Methods for details), a machine learning technique used in bioinformatics research to predict sample type with a small number of features (86, 87). Using SVM-RFE, 20 predictive features were successfully selected for each CRISPR and 16S rRNA (Figure 11). For STEQ CRISPR, 20 spacers achieved 95.2% accuracy for personal identification, which was the highest performance in our datasets. Although 20 spacers from STTH achieved 85.0% accuracy, the accuracy rates obtained using selected features of MILU or 16S rRNA were 47.1 % and 52.6%, respectively. The relative abundance heatmap of the 20 spacers showed that the presence of spacers in streptococcal CRISPRs had a distinct and specific pattern for each individual (Figure 11B, C), whereas MILU spacers did not (Figure 11D). These individual specific spacers likely contributed to the higher accuracy. Therefore, the spacer pattern of even a single CRISPR array, such as STEQ CRISPR, may sufficiently discriminate personal microbiome signatures for future personal identification.

Discussion

CRISPR typing has been used to discriminate bacterial species at the strain-level (33-37, 39). This study proposed metagenomic CRISPR typing as a new personal microbiome typing method, instead of conventional 16S rRNA V1-2/V4 sequencing (11, 16, 88). In chapter I, putative CRISPRs shared between the skin microbiomes of different individuals were identified by metagenomic reconstruction of CRISPR arrays. In this chapter, the distribution of spacers in the skin microbiome was further examined by amplicon sequencing, revealing that CRISPR diversity was highly person-specific

regardless of skin site, unlike the skin microbiome structure assessed by 16S rRNA V1-3 sequencing (Figure 6, Figure 9, and Figure 10).

Spacer sequences in the human skin microbiome were amplicon sequenced to identify the CRISPRs that are shared by multiple individuals and are highly individual-specific. Although the number of participants was limited, the spacers were successfully amplified from all five individuals, supporting the notion that putative CRISPRs in the metagenomic datasets were conserved among individuals. As hypothesized, the diversity of the STEQ and STTH spacers were highly distinct, whereas major spacers of the MILU CRISPR were shared between individuals (Figure 6 and Figure 11), indicating that STEQ and STTH are good candidates for future personal identification markers. Unexpectedly, even major spacers were shared between samples collected at different timepoints from the same individual, suggesting that spacer diversity was stable over two months (Figure 6 and Figure 11). Yang et al. demonstrated that 16S rRNA V1-2 sequencing is not sufficient to discriminate personal microbiomes (27). This study also supports that skin microbiome composition measured by 16S RNA sequencing is insufficient for personal discrimination (Figure 10 and Figure 11). In contrast to bacterial communities, membership of bacteriophage communities was mostly individual-specific (32). The membership of CRISPR spacers also reflects this individual specificity of bacteriophages, which likely contributes to high inter-individual dissimilarity. This speculation was supported by the observation of phage derived sequences in the predictive spacers of streptococcal CRISPRs (Figure 11B and C). Thus, polymorphisms of CRISPR spacers, rather than conventional 16S rRNA, may be useful for discriminating between suspects in criminal investigations.

Two CRISPR arrays, STEQ and STTH, were derived from *Streptococcus* species,

which are also known as *S. thermophilus* CRISPR-Cas9 system in CRISPR1 and CRISPR3 loci, respectively. The CRISPR loci of *S. thermophilus* have been intensively studied in both evolutionary and genome editing perspectives (63). The CRISPR arrays of *S. thermophilus* are classified into four CRISPRs: CRISPR1 (STEQ), CRISPR2, CRISPR3 (STTH), and CRISPR4. Paez-Espino et al. revealed that spacer acquisition ability differs between them, as CRISPR1 integrated new spacers at a high frequency while CRISPR2 and CRISPR4 acquired no spacers in 232 days (89). In contrast to *Streptococcus* CRISPRs, which integrate a new spacer frequently, the spacer integration of MILU spacers, which is derived from *M. luteus*, has not been investigated but the *cas* genes of *M. luteus* strain SGAir0127 were classified into type I CRISPR-Cas system by CRISPRCasFinder. Although *M. luteus* is also a major member of the skin microbiome (12), the only available MILU CRISPR array in the NCBI database was that of strain SGAir0127, a strain isolated from indoor air samples (90). Though the nature of MILU CRISPR is unknown, it was speculated that it lost its spacer acquisition ability during its evolution.

This study focused on the diversity of spacers on a community level and did not comprehensively evaluate the temporal changes of single CRISPR arrays. The spacer acquisition ability of the streptococcal CRISPR-Cas system, which may contribute to the high inter-individual spacer dissimilarity observed in this study, should be further assessed using other technologies, such as CRISPR locus PCR with the primer pair flanking an entire CRISPR region (38, 40-42). Nonetheless, partial or entire CRISPR arrays were successfully reconstructed from amplicon reads (Figure 7), suggesting that amplicon sequencing may be utilized to phylogenetically analyze CRISPR loci of interest, while technical optimizations, such as utilizing size selection during library construction

and/or other sequence chemistries such as the paired-end sequencing of 600 cycles, will improve the shortcomings of the methodology on the CRISPR locus reconstruction analysis presented in this study.

Although the resolution of CRISPR typing is sufficiently high for strain discrimination, other typing methods such as multilocus sequence typing, variable number tandem repeats typing (VNTR-typing) or 16S ribotyping can also be used to discriminate specific bacterial strains (65, 91-93). Recent studies also utilized such high-resolution typing methods in metagenomic contexts to characterize the composition of a specific species or bacterial communities (27, 94, 95). High interpersonal diversity of spacer sequences measured in this study demonstrate its potential of application in future forensic cases; however, it should be compared with other typing methods, such as *C. acnes* ribotyping. Importantly, the diversity of CRISPR arrays were similar between different skin sites from the same individual (Figure 6, Figure 9, and Figure 11), suggesting that CRISPR spacers are shared among the skin sites. This spatial stability of CRISPR typing may be an advantage against *C. acnes* ribotyping, in which only palm samples were analyzed (27).

Though the results presented in this chapter suggest that spacer diversity is highly personalized and useful as a new marker for personal identification, some aspects regarding the real-life applicability of this approach require further analysis. For example, the approach presented in this study was not used to test the microbiome on the surface of an object, such as a phone, a keyboard or shoes, touched by a person of interest (suspect/victim) (11, 16, 27, 96). Another limitation of this study is the small sample size. Although it was demonstrated that spacer diversity was sufficiently individual-specific, the application in forensic cases requires population-wide validation to assess their

validity and reliability. In addition, the performance in terms of personal identification was evaluated only for comparing the CRISPR-based and 16S rRNA-based approaches. Therefore, the score obtained in this study does not predict the performance of CRISPR typing on personal identification in a population, although 95.2 % accuracy was obtained using STEQ as a marker. By confirming the accuracy of CRISPR-based personal identification using mock forensic samples and conducting population-wide surveys in future studies, CRISPR-based microbiome analysis can be applied in actual forensic cases.

Although further developmental validation is required, the highly personalized diversity of CRISPR spacers in the human skin microbiome is an attractive tool for personal identification. Though the viral community is more personalized than the bacterial community (31, 32, 50), the lack of a common marker for characterizing the virome complicates its use for personal identification. In this study, the spacer sequences of CRISPRs were utilized as a proxy for the virome. This approach may also be utilized to better understand the ecology of the skin microbiome and explore the complex relationship between the human skin microbiome and diseases such as atopic dermatitis (97). The present study suggests that the diversity of the CRISPR spacer sequences in the skin microbiome can be used as a new marker for personal identification.

Summary

Streptococcal CRISPRs can be used as a genetic marker to discriminate personal microbiome, however, little is known about diversity of Streptococcal CRISPRs in human skin microbiome. In this chapter, streptococcal CRISPRs were comprehensively sequenced by PCR-based sequencing. A total of 1223462 reads from three CRISPR arrays revealed that spacers in the skin microbiome were highly personalized, and conserved repeats were commonly shared between individuals. These individual specificities observed using CRISPR typing were confirmed by comparing the CRISPR diversity to microbiome diversity assessed using 16S rRNA amplicon sequencing. Furthermore, it was tested whether a set of informative spacers can discriminate the personal microbiome by utilizing SVM-RFE, showing that CRISPR typing achieved 95.2 % accuracy, whereas 16S rRNA sequencing only achieved 52.6 %. These results suggest that sequencing streptococcal CRISPRs in the skin microbiome may be a more powerful approach for personal identification and ecological studies compared to conventional 16S rRNA sequencing.

Table 6 Metadata for all samples collected in this chapter.

Sample ID	Participant ID	Sampling Date	Gender	Age Range	Systemic antibiotics usage in 6 months	Steroids usage in 6 months	Skin care products usage in 24 hours	Hand wash in 2 hours	Site-Symmetry	Site characteristic
MCP000	P04	2019-05-30	Male	20s	No	No	No	Yes	Hp-L	Dry
MCP001	P01	2019-07-02	Male	30s	No	Yes	Yes	Yes	Hp-R	Dry
MCP003	P02	2019-07-02	Female	20s	Yes	No	Yes	Yes	Hp-R	Dry
MCP005	P03	2019-07-03	Female	40s	No	No	No	Yes	Hp-R	Dry
MCP008	P04	2019-08-07	Male	20s	No	No	No	Yes	Hp-R	Dry
MCP010	P04	2019-08-07	Male	20s	No	No	No	Yes	Ac-R	Wet
MCP012	P04	2019-08-07	Male	20s	No	No	No	Yes	Ra-R	Sebaceous
MCP013	EC	2019-08-07							EC	EC
MCP014	P01	2019-09-02	Male	30s	No	Yes	No	No	Hp-L	Dry
MCP016	P01	2019-09-02	Male	30s	No	Yes	No	No	Ac-L	Wet
MCP018	P01	2019-09-02	Male	30s	No	Yes	No	No	Ra-L	Sebaceous
MCP020	P03	2019-09-02	Female	40s	No	No	Yes	Yes	Hp-L	Dry
MCP023	P03	2019-09-02	Female	40s	No	No	Yes	Yes	Ac-R	Wet
MCP024	P03	2019-09-02	Female	40s	No	No	Yes	Yes	Ra-L	Sebaceous
MCP027	P02	2019-09-12	Female	20s	No	No	No	Yes	Hp-R	Dry
MCP028	P02	2019-09-12	Female	20s	No	No	No	Yes	Ac-L	Wet
MCP031	P02	2019-09-12	Female	20s	No	No	No	Yes	Ra-R	Sebaceous
MCP039	P05	2019-12-12	Male	30s	Yes	No	No	Yes	Hp-R	Dry
MCP041	P05	2019-12-12	Male	30s	Yes	No	No	Yes	Ac-R	Wet
MCP043	P05	2019-12-12	Male	30s	Yes	No	No	Yes	Ra-R	Sebaceous
MCP050	EC	2020-01-20							EC	EC
MCP057	P05	2020-03-09	Male	30s	No	No	Yes	Yes	Hp-L	Dry
MCP060	P05	2020-03-09	Male	30s	No	No	Yes	Yes	Ac-R	Wet
MCP061	P05	2020-03-09	Male	30s	No	No	Yes	Yes	Ra-L	Sebaceous

Table 7 Primer sequences used in this study.

Name	Sequence 5' > 3'	Ref
iIF-STEQ_DR_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGAAACAACACAGCTCTAAAAC	
iIR-STEQ_DR_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTTGTTTCGAATGGTTCCAAAAC	
iIF-STTH_DR_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTCTCAAGATTTAAGTAACTGTACAAC	
iIR-STTH_DR_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGTTACTTAAATCTTGAGAGTACAAAAAC	
iIF-MILU_DR_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGCGGACTTCGTTGAGGC	
iIR-MILU_DR_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCCCCACAGGAGCGGAGAC	
iIF-27F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGTTTGATCMTGGCTCAG	
iIR-534R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTACCGCGGCTGCTGG	
Nadkarni_F	TCCTACGGGAGGCAGCAGT	(98)
Nadkarni_R	GGACTACCAGGGTATCTAATCCTGTT	(98)

Table 8 Sequencing statistics for spacer sequencing.

Repeat ID	Sample ID	Subject ID	Number of reads	Average length	Number of reads after trimming	% reads kept after trimming	Average length after trimming
STEQ	MCP000	P04	22476	97.8	22017	97.96%	30
STEQ	MCP001	P01	45298	94.9	43661	96.39%	30
STEQ	MCP003	P02	47090	97.5	43853	93.13%	30
STEQ	MCP005	P03	39854	95.3	38895	97.59%	30
STEQ	MCP008	P04	41256	151	40609	98.43%	30.1
STEQ	MCP010	P04	46400	151	45093	97.18%	30
STEQ	MCP012	P04	57552	151	56564	98.28%	30
STEQ	MCP013	EC	36462	151	33436	91.70%	29.6
STEQ	MCP014	P01	104016	89.5	101880	97.95%	30
STEQ	MCP016	P01	66056	96.5	64536	97.70%	30.0
STEQ	MCP018	P01	6864	93.9	5161	75.19%	30
STEQ	MCP020	P03	55094	88.5	53462	97.04%	30.1
STEQ	MCP023	P03	68372	95.1	66489	97.25%	29.9
STEQ	MCP024	P03	15316	92.4	13636	89.03%	30.2
STEQ	MCP027	P02	23984	86.8	23039	96.06%	30
STEQ	MCP028	P02	98700	97.3	79021	80.06%	29.1
STEQ	MCP031	P02	28486	103.5	19084	66.99%	29.6
STEQ	MCP039	P05	127392	93.3	125279	98.34%	30.0
STEQ	MCP041	P05	97342	97.9	91871	94.38%	29.8
STEQ	MCP043	P05	75890	95.3	64638	85.17%	28.9
STEQ	MCP050	EC	12282	126.4	3738	30.43%	29.2
STEQ	MCP057	P05	77872	92.0	75546	97.01%	30.0
STEQ	MCP060	P05	72374	93.8	69941	96.64%	30.0
STEQ	MCP061	P05	84210	95.9	79993	94.99%	30.0
STTH	MCP000	P04	47982	98.8	41361	86.20%	30
STTH	MCP001	P01	47342	100.8	41724	88.13%	29.9
STTH	MCP003	P02	32664	95.6	26826	82.13%	30
STTH	MCP005	P03	40444	93	32075	79.31%	29.9
STTH	MCP008	P04	47884	151	46391	96.88%	30
STTH	MCP010	P04	46550	151	44784	96.21%	30
STTH	MCP012	P04	40254	151	37945	94.26%	30
STTH	MCP013	EC	28252	151	4071	14.41%	27.5
STTH	MCP014	P01	32286	99.9	31158	96.51%	29.8
STTH	MCP016	P01	81590	96.1	74543	91.36%	29.8
STTH	MCP018	P01	63556	73.5	35228	55.43%	29.9
STTH	MCP020	P03	37676	96.4	35507	94.24%	29.9
STTH	MCP023	P03	3986	51.9	855	21.45%	29.8
STTH	MCP024	P03	114390	87.4	88743	77.58%	29.9
STTH	MCP027	P02	149238	96.2	141020	94.49%	29.9
STTH	MCP028	P02	101118	55.2	27356	27.05%	29.8
STTH	MCP031	P02	37242	44.4	1888	5.07%	28.4
STTH	MCP039	P05	73466	97.3	68699	93.51%	29.9
STTH	MCP041	P05	80116	83.1	55445	69.21%	29.8
STTH	MCP043	P05	210646	85.2	154578	73.38%	29.8
STTH	MCP050	EC	25018	49.7	2883	11.52%	29.6
STTH	MCP057	P05	73634	80.0	51427	69.84%	29.9
STTH	MCP060	P05	135314	93.7	120872	89.33%	29.8
STTH	MCP061	P05	70174	56.3	18657	26.59%	29.8
MILU	MCP000	P04	46426	100.6	45206	97.37%	35.8
MILU	MCP001	P01	44944	103.3	43641	97.10%	36
MILU	MCP003	P02	3356	106.3	2974	88.62%	36.5
MILU	MCP005	P03	15014	97.8	14365	95.68%	35.9
MILU	MCP008	P04	74810	151	71252	95.24%	35.8
MILU	MCP010	P04	86940	151	83123	95.61%	35.8
MILU	MCP012	P04	76306	151	71612	93.85%	36.2
MILU	MCP013	EC	15626	151	153	0.98%	31.1
MILU	MCP014	P01	40720	94.3	39216	96.31%	36
MILU	MCP016	P01	63274	98.2	59881	94.64%	36.3
MILU	MCP018	P01	60962	86.7	57163	93.77%	35.9
MILU	MCP020	P03	31122	92.9	30166	96.93%	35.8
MILU	MCP023	P03	2508	82.8	1093	43.58%	33.7
MILU	MCP024	P03	1950	101.2	1081	55.44%	33.8
MILU	MCP027	P02	47014	91.2	45020	95.76%	35.9
MILU	MCP028	P02	40672	96.1	36696	90.22%	35.6
MILU	MCP031	P02	47460	93.4	45886	96.68%	35.4
MILU	MCP039	P05	43154	96.2	40357	93.52%	35.4
MILU	MCP041	P05	7598	104.0	6562	86.36%	33.2
MILU	MCP043	P05	20260	95.1	17896	88.33%	37.0
MILU	MCP050	EC	594	126.2	25	4.21%	32.9
MILU	MCP057	P05	32050	81.7	30996	96.71%	36.5
MILU	MCP060	P05	19232	99.0	18085	94.04%	38.6
MILU	MCP061	P05	1892	147.4	60	3.17%	32.8

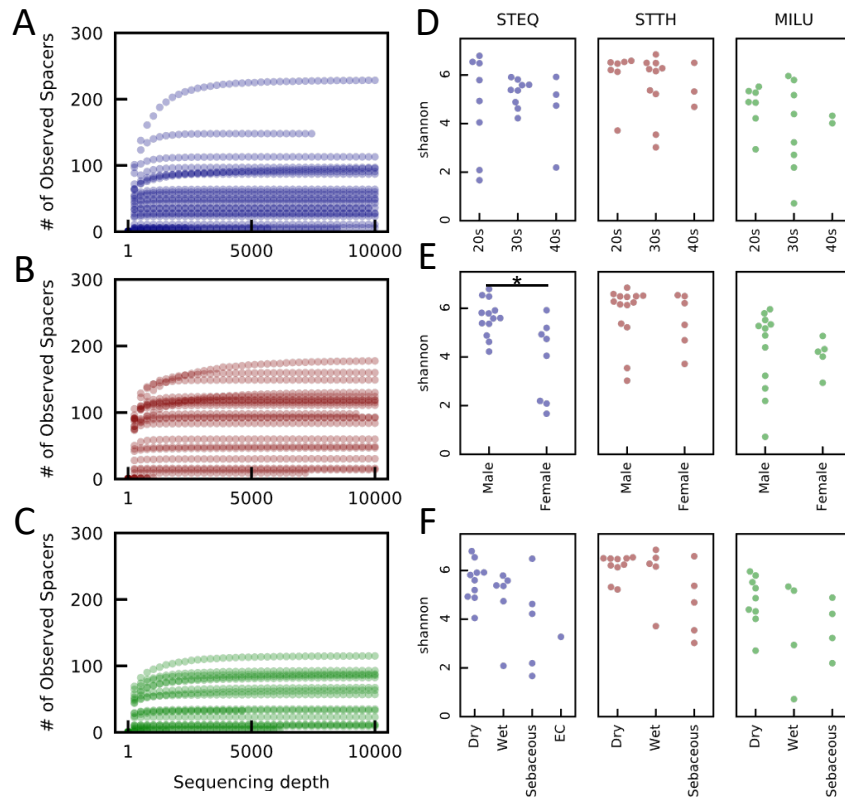


Figure 4 Alpha diversity of CRISPR spacer sequences in the skin microbiome.

(A-C) Rarefaction curves based on the ASVs of spacer sequences were calculated using the q2-diversity plug-in in QIIME 2 with 100 iterations and a maximum sequence depth of 10000. STEQ (A), STTH (B), and MILU (C) spacers. (D-F) The Shannon diversity index of each sample was calculated using the q2-diversity plug-in in QIIME 2. The Shannon indices of each sample group were compared using the Kruskal–Wallis test ($*p < 0.05$). Comparisons in terms of age range (D), gender (E), and site characteristics (F).

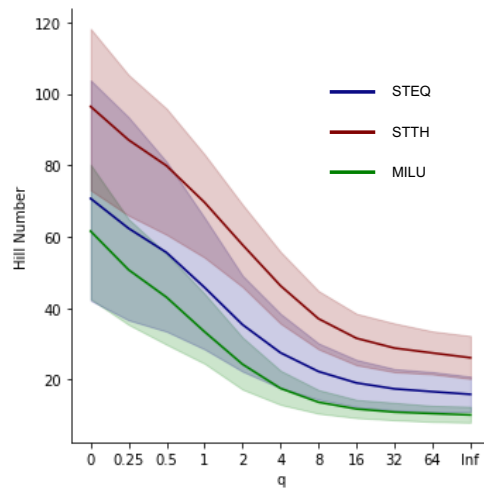


Figure 5 Hill numbers plotted as a function of the parameter.

The Hill numbers were calculated using the library, “vegan” in R.

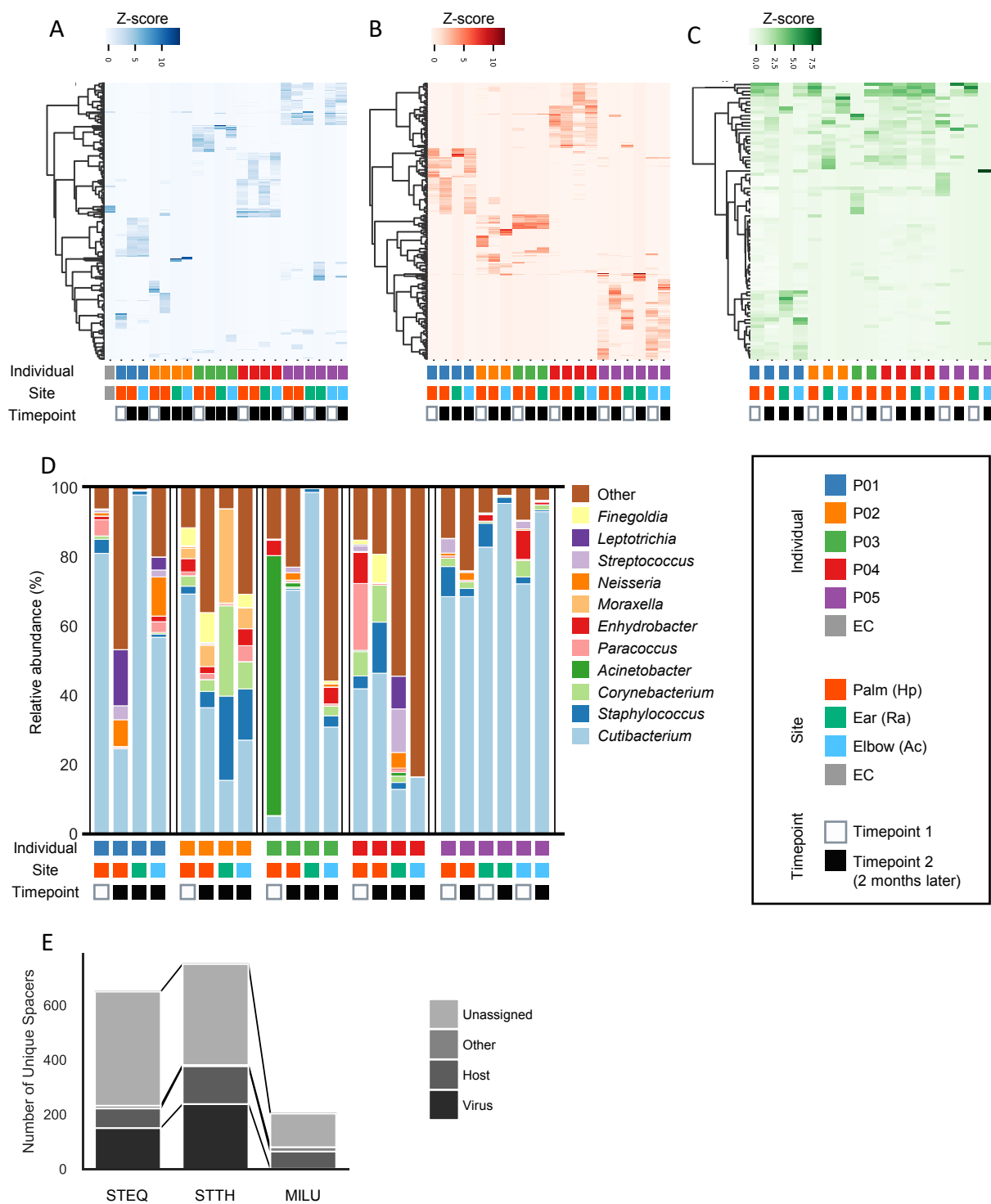


Figure 6 The pattern of CRISPR spacer and that of microbiome measured by 16S RNA in the skin microbiome.

(Continued)

(A-C) A heatmap of the CRISPR spacers in the skin microbiome. Each row represents a unique spacer sequence ASV and the columns represent samples from five individuals. Each row was clustered by the Wards method implemented in the Seaborn package. Each column was standardized by the z-score. The ASVs that were observed equal or more than 100 times in total were used to generate the heatmaps. STEQ (A), STTH (B), and MILU (C) spacers. (D) Skin microbiome profile measured by 16S rRNA V1-3 sequencing. The composition of the top 11 genera is shown. (A–D) Boxes below the graph indicate sampled individual, site, and timepoint. (E) Numbers of unique spacer sequences and classification of the sequences were compared by CRISPRs.

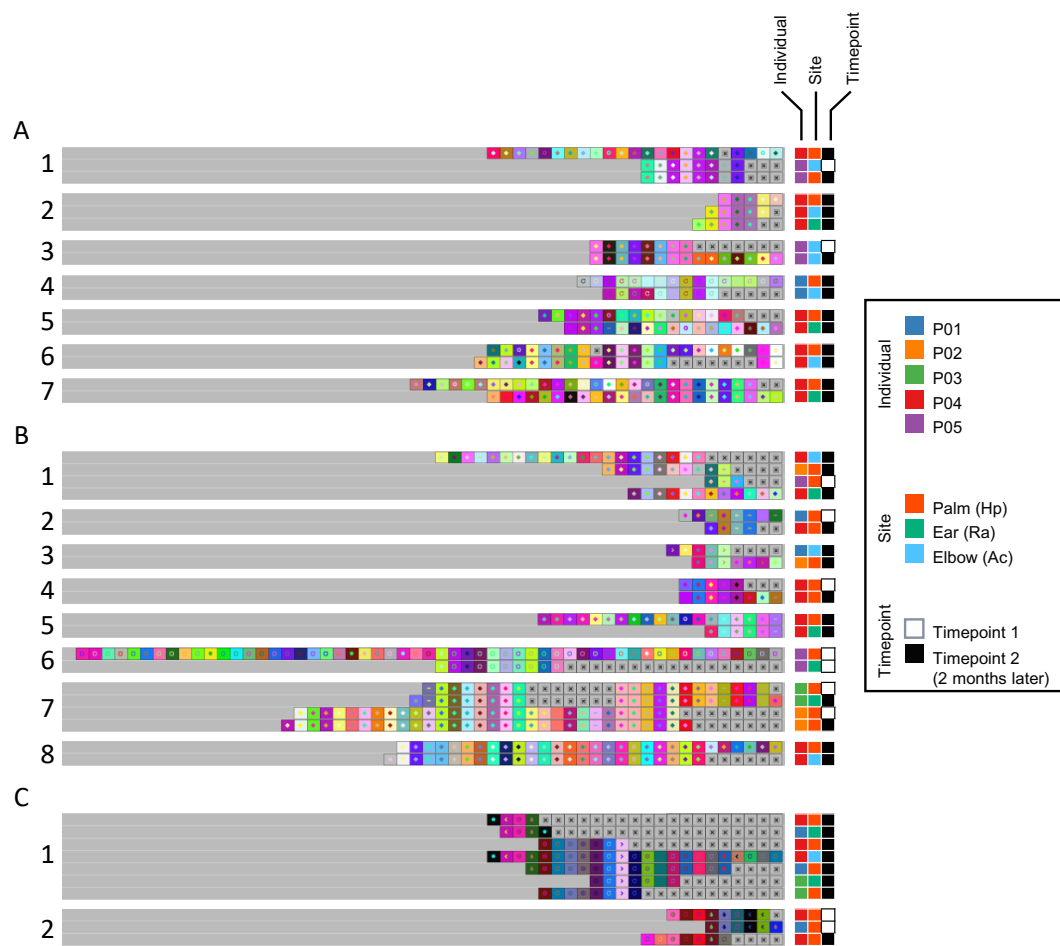


Figure 7 Comparison of putative CRISPR arrays reconstructed from amplicon sequencing.

(A–C) The STEQ (A), STTH (B), and MILU (C) CRISPR arrays were reconstructed from spacer amplicons and analyzed by CRISPR Visualizer (62). CRISPR arrays sharing spacers were grouped together. Each colored box represents a spacer. Grey boxes with “x” represent missing spacers. Note that similar colors do not indicate similar sequences.

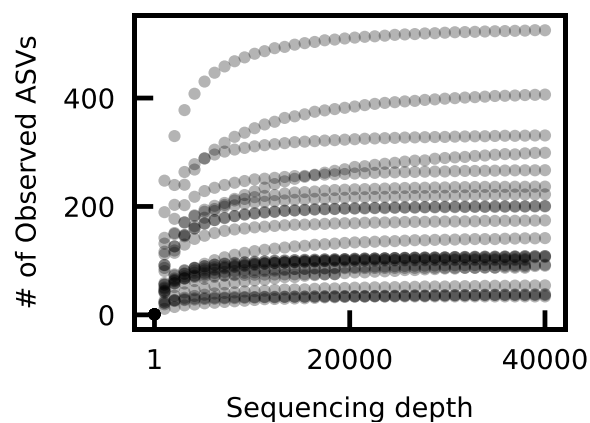


Figure 8 Rarefaction curve analysis of 16S rRNA V1-3 sequencing data.

Rarefaction curves based on the ASVs of 16S rRNA V1-3 sequences were calculated using the q2-diversity plug-in in QIIME 2 with 100 iterations and a maximum sequence depth of 40000.

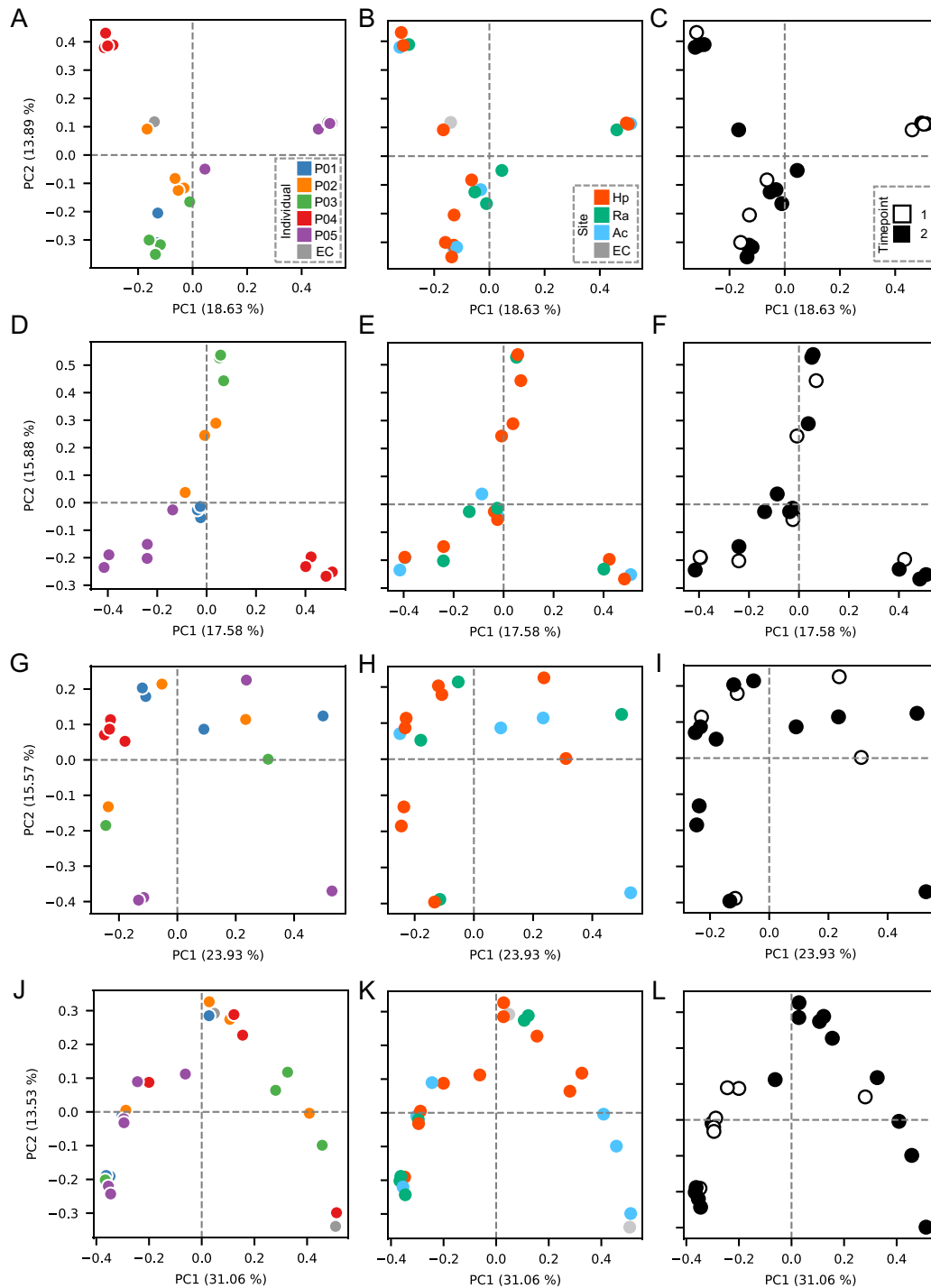


Figure 9 Principal coordinate analysis of the spacers and 16S rRNA.

(A–L) PCoA plots of the STEQ (A–C), STTH (D–F), and MILU (G–I) spacer sequences as well as the 16S rRNA V1-3 sequences (J–L). The Bray-Curtis dissimilarity index was used as a metric for PCoA. Each datapoint is colored according to the individual (A, D, G, and J), skin site (B, E, H, and K), or timepoint (C, F, I, and L).

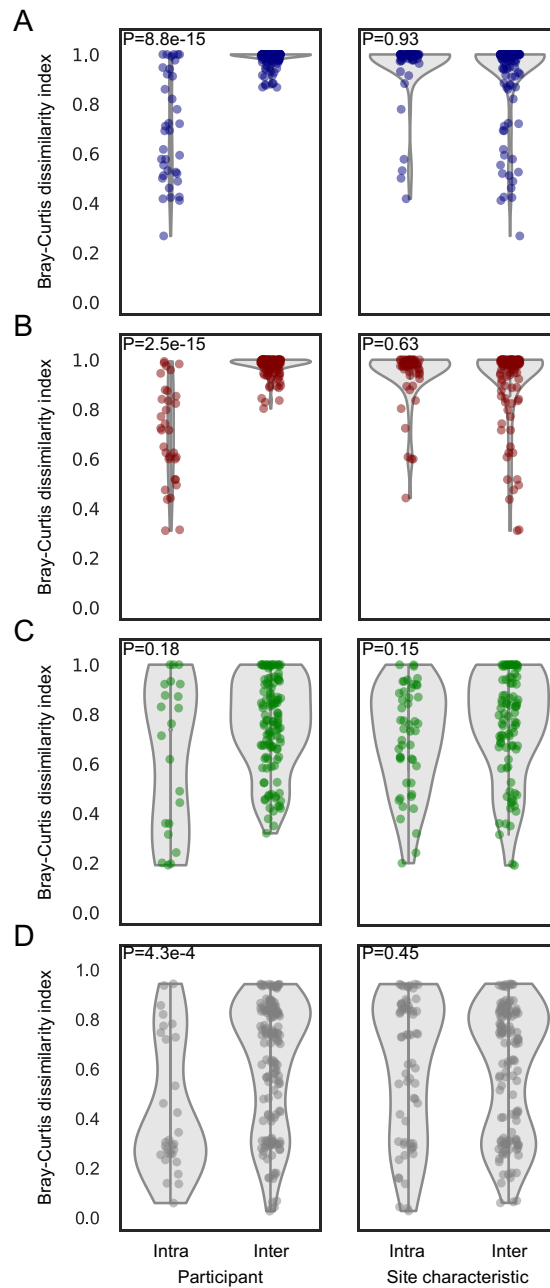


Figure 10 CRISPR spacers are individual-specific rather than site-specific.

(A–D) Violin plots of the Bray-Curtis dissimilarity indices within (“intra”) or between (“inter”) individuals/hosts site of STEQ (A), STTH (B), MILU (C) and 16S rRNA V1-3 (D). The Bray-Curtis dissimilarity indices were compared using the Mann–Whitney U test.

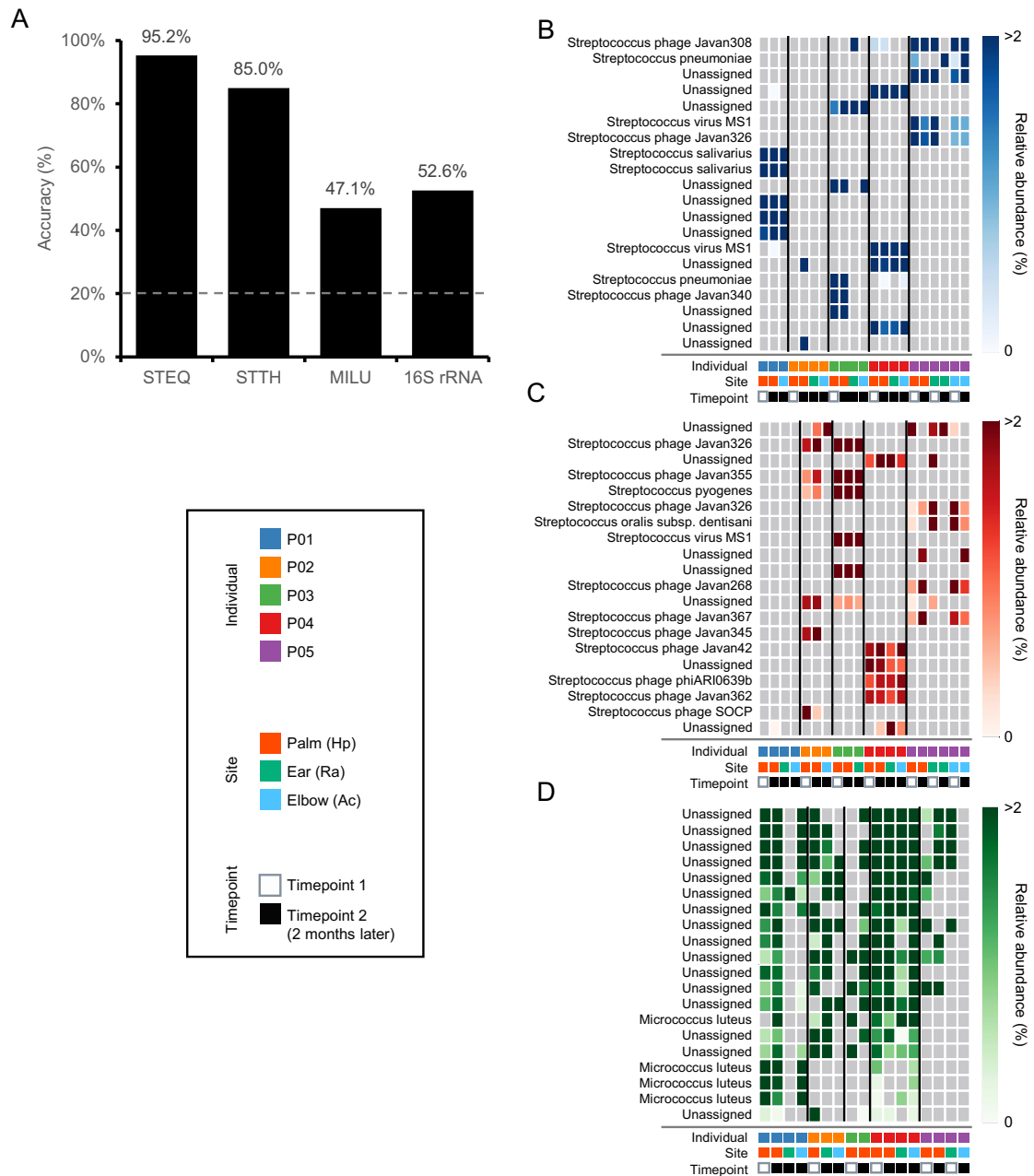


Figure 11 Identification of the predictive spacer sequences for personal identification.

(A) Comparison of accuracy results using STEQ, STTH and MILU CRISPRs and 16S rRNA data. The accuracy results were obtained using 20 features selected by SVM-RFE. After feature selection, SVM with leave-one-out cross-validation was performed to classify the origins of the samples. Baseline accuracy is indicated by grey dashed line. (B–D) Heatmap of the relative abundance of the 20 selected features for STEQ (B), STTH (C), and MILU (D). Data points with absent features are greyed out. The annotation of each spacer sequence was based on the BLAST search.

CHAPTER III

Optimization of microbial DNA extraction from human skin samples for CRISPR typing

Introduction

Short tandem repeat (STR), which is also known as microsatellite or simple sequence repeat, is a repetitive structure of genomic DNA containing array of short repeat units which are typically several base pairs in length (99). DNA profiling using STR typing is the current standard method for forensic human identification, an essential procedure in criminal investigations. However, STR typing of samples touched by a person of interest is difficult because of the low levels of DNA (18, 100). For example, Daly et al. recovered and quantified trace DNA from glass objects held by volunteers to evaluate the transfer of DNA from human hands (18). They obtained less than 0.65 ng of human DNA from 74% of the samples and no profile from 19/40 samples using the AmpFISTR® SGM Plus™ Amplification Kit (Applied Biosystems) on an ABI PRISM 3130 Genetic Analyzer (Applied Biosystems), indicating the difficulty in DNA profiling using human DNA recovered from touched objects. Therefore, alternative genetic markers may be used to complement such inconclusive STR profiling.

To complement STR profiling, previous reports have proposed the microbiome as a new forensic genetic marker for human identification (11, 15, 23, 25, 27, 88, 100, 101). In chapter II, spacers of *Streptococcus* CRISPRs were sequenced by amplifying the spacer sequences using primer pairs targeting the repeats to characterize the human skin

microbiome for forensic individual identification. The sequencing of two *Streptococcus* CRISPRs achieved higher accuracy in human identification than conventional 16S rRNA sequencing; however, the validity and reliability in forensic casework could be assessed further by optimizing each technical experimental protocol step.

DNA extraction, an essential step in the processing of a skin microbiome sample, severely affects the results of downstream analysis because the DNA extraction protocol can alter DNA recovery efficiency or microbial composition estimation accuracy (102, 103). Thus, we need to choose an appropriate DNA extraction protocol for microbial genomic DNA extraction from human skin samples (102, 104).

Forensic researchers have made significant efforts to automate nucleic acid extraction from a variety of crime scene samples to handle an increasing number of samples. EZ1 Advanced XL (Qiagen, Aarhus, Denmark), along with the EZ1 DNA Investigator Kit, is a representative convenient automated nucleic acid extraction platform that is used in many forensic laboratories (23, 105-107). As this platform is widely used in forensic testing, forensic analysts can use extracted DNA for both human and microbial DNA analysis if the platform is also available for microbiome analysis; however, its use in microbiome studies is limited. While Nishi et al. utilized this platform to investigate the human skin microbiome (23), it is still debatable whether the genomic DNA extracted using the EZ1 platform for forensic DNA typing is also suitable for analyzing the microbiome.

In this chapter, the EZ1 platform was compared with a widely used column-based manual DNA extraction kit, the Qiagen DNeasy PowerSoil DNA isolation kit, which has been used previously, to benchmark the suitability of the EZ1 platform for microbiome studies (2, 11, 70). The results of chapter III, which expand the results of

chapter II by analyzing 40 additional samples, suggest that these two DNA extraction protocols are comparable in terms of both quantity and quality of extracted microbial DNA.

Materials and methods

Sample collection

Sample collection procedures have been described in chapter II. Five healthy adult human volunteers were recruited for this study, and skin swabs were collected for spacer sequencing. Table 6 and Table 9 summarizes the detailed sample information. Skin surfaces (approximately 5×5 cm) were swabbed with cotton-tipped swabs moistened with water. Swabs were sampled from the Hp, Ra and Ac, representing dry, sebaceous, and moist sites, respectively (12). For the initial sampling (time point 1), only the Hp sample was obtained, except for one participant (P05). Two to three months later (time point 2), three skin sites (Hp, Ac, and Ra) were sampled from all participants. Approximately 6 months after the initial sampling (time point 3), the three skin sites were sampled from three participants, P01, P03, and P04.

DNA extraction

Genomic DNA was extracted from swabs using the EZ1 DNA Investigator Kit (Qiagen; EZ1) or the DNeasy PowerSoil Kit (Qiagen, Aarhus, Denmark; PS). For the EZ1 kit, DNA was extracted using the EZ1 DNA Investigator Handbook (April 2009). Briefly, frozen cotton tips were transferred into 2-ml tubes containing 290 μ l of 2-fold diluted Buffer G2. After the addition of 10 μ l of proteinase K, the tubes were incubated at 56 °C for 15 min and at 95 °C for 5 min. Then, up to 200 μ l of the solution was transferred into

a new tube and subjected to “Trace Protocol” on EZ1 Advanced XL (Qiagen). We did not make any modifications, such as lysozyme treatment or bead beating, in the EZ1 protocol. Genomic DNA was eluted using 50 µl of water. For the PS kit, DNA was extracted as described in chapter II. The samples analyzed with the PS protocol included both newly obtained skin swab samples and those obtained in chapter II. Specifically, descriptions of the DNA extraction methods and individual sample sources newly analyzed are provided in Table 9.

Quantification of 16S rRNA gene copy number

To quantify bacterial DNA, 16S rRNA gene quantification was performed through qPCR using SYBR Premix Ex Taq (TaKaRa Bio, Shiga, Japan) or TB Green Premix Ex Taq (Tli RNase Plus) (Takara), and QuantiStudio 5 (Thermo Fisher Scientific, Waltham, MA) with the primer pair targeting the bacterial 16S rRNA gene, Nadkarni_F and Nadkarni_R (98). The primer sequences are summarized in Table 7. The standard curve was generated by serially diluting genomic DNA of *C. acnes* JCM6425 from 1 ng/µl to 0.1 pg/µl. Each reaction was performed in a reaction volume of 10 µl containing 5 µl of 2 × SYBR Premix Ex Taq or 2 × TB Green Premix Ex Taq (Tli RNase Plus), 0.4 µM of each primer, 1 × ROX II dye, and 2 µl of genomic DNA. An NTC was used for each reaction. The following conditions were used for amplification: 30 s at 95 °C, 40 cycles of denaturation at 95 °C for 5 s, annealing and extension at 60 °C for 34 s, followed by melting curve analysis from 60 °C to 95 °C. Each qPCR was performed in triplicate. Note that all of the calculated DNA concentrations for EC and NTC samples were below the lower limit of the standard curve, 0.1 pg/µl.

Amplification and sequencing of spacers

Spacer sequencing was performed as described in chapter II. In this chapter, two streptococcal CRISPRs, CRISPR3 and CRISPR1 which are referred to as STEQ and STTH in chapter II (38), were analyzed using two primer pairs (Table 7).

Statistical analysis

The Mann-Whitney U test was performed using “stats.mannwhitneyu” function of the python library SciPy (version 1.3.0) to compare DNA concentrations extracted from skin swab samples.

For spacer analysis, statistical analysis was performed as described in chapter II, with minor modifications. The parameters and versions of the software were the same as those in chapter II, unless otherwise stated. Initial quality control and ASV table construction were performed using the CLC Genomics Workbench (Qiagen) and q2-dada2 plug-in of QIIME 2 v2020.2 (54, 73). Then, each sample was rarefied to 3158 reads for normalization, and samples with less than 3158 reads were omitted from the calculation. The EC and NTC samples were removed from the ASV table in the following analysis. The ASV table was then standardized by calculating the z-score and incorporated into R (version 4.0.2). A heatmap was generated from the table using the ComplexHeatmap library (version 2.6.2) to visualize spacer diversity (108). To perform individual prediction, we used a SVM-REF (77). SVM-RFE was implemented using the Python library “Scikit-learn” (78). Using SVM-RFE with a linear kernel, 20 ASVs were selected from each dataset, including samples grouped by extraction protocols or by combining the two datasets. Then, using the selected 20 features, the individuals were predicted using linear-kernel SVM via leave-one-out cross-validation to measure

predictive accuracy.

Results

Comparison of DNA extraction protocol based on amount of microbial DNA

To evaluate the microbial DNA yields of different DNA extraction protocols, microbial genomic DNA were extracted from a total of 31 skin swabs using the EZ1 DNA Investigator Kit in combination with EZ1 Advanced XL, which is commonly used to obtain human genomic DNA from forensic specimens. The results were compared with those of 33 skin swabs from which microbial DNA was extracted using the PS protocol, which is commonly used in microbiome studies (2, 11, 70). The DNA recovery of both protocols were evaluated by quantifying the amount of DNA using qPCR with universal primers targeting bacterial 16S rRNA. As shown in Figure 12, the median DNA yield of the EZ1 protocol was 330.5 pg (IQR: 68.2-725.6), while that of the PS protocol was 196.4 pg (IQR: 75.5-1038.3). There were no significant differences between the two protocols. Thus, these data suggest that the performance of the EZ protocol in terms of microbial DNA yield is sufficient for microbiome studies.

Comparison of DNA extraction protocol based on spacer diversity analysis

While the results of DNA quantification supported efficient DNA recovery with the EZ1 protocol, the DNA extraction method may also alter the results of the microbiome analysis (102). It is evaluated whether the DNA extraction protocol altered the composition of CRISPRs, by comparing the spacer diversity of the skin swab samples treated with different DNA extraction protocols. As shown in Figure 13, the spacer diversity patterns were distinct among the participants, in accordance with the previous

report revealing highly personalized spacer diversity (50). For example, spacer cluster 1 on CRISPR3 was specifically and stably observed in participant P03, whereas cluster 3 of CRISPR3 corresponded to participant P05. Conversely, the signatures were not altered by DNA extraction protocols for both *Streptococcus* CRISPRs, while some discrepancies were observed between sample pairs, such as wet site samples collected from participant P03 at time point 2 for CRISPR3 analysis. All sample pairs analyzed with different DNA extraction protocols were derived from opposite sides of the same skin site (for example, left Hp and right Hp). Thus, anatomical differences in the skin site may also cause discrepancies in the spacer patterns. These results suggest that the difference in DNA extraction protocol may not alter the results of spacer diversity analysis, but it is unknown how many CRISPR arrays are required to obtain stable results through CRISPR analysis.

DNA extraction protocol does not influence predictive accuracy of spacer diversity analysis

In chapter II, metagenomic CRISPR typing was utilized to predict individuals from spacer diversities using a machine learning approach, SVM-RFE. Individual prediction models were built for EZ1 and PS samples to evaluate the influence of DNA extraction protocols on the accuracy of human identification. The accuracies of individual predictions from DNA extracted using EZ1 were comparable to those of the PS protocol (Table 10), suggesting that the influence of DNA extraction protocols can be ignored in individual predictions. Moreover, the highest accuracies were obtained by combining both sample sets, probably because roughly double samples were used as training data during the construction of each predictive model.

Discussion

In this chapter, two DNA extraction protocols, EZ1 and PS, were compared to optimize the metagenomic CRISPR typing method for forensic applications. While a previous report described the use of the EZ1 platform in microbial community analysis via T-RFLP analysis (23), they only compared the efficiency of DNA yield. Thus, to our knowledge, this is the first report showing both the efficiency of the EZ1 protocol based on DNA yield and the influence of the protocol on community structure analysis and individual prediction accuracy. The results indicated that the microbial DNA yields of the automated EZ1 protocol were comparable to those obtained using the manual PS protocol. Furthermore, the differences in the DNA extraction methods did not affect the profiles of metagenomic CRISPR typing or the prediction accuracy of individual identification. These results suggest that the EZ1 protocol could also be used for microbiome research, including metagenomic CRISPR typing, while further validation, such as analysis of mock casework samples, should be performed before the application of microbiome analysis for forensic case studies.

Human skin microbiome samples are regarded as low-biomass; therefore, the selection of DNA extraction methods can skew the microbial community analysis results (72). The DNA yields in both the previous study (23) and this study were typically below 10 ng, which was far less than the DNA yields of fecal microbiome specimens, which are typically above 100 ng (109, 110). Generally, low DNA yield may lead to unstable results in microbiome analysis, due to both contamination and stochastic effects (80, 111). Thus, such characteristics of skin specimens might cause partial or whole spacer loss in some samples analyzed in this study (Figure 13), suggesting the importance of DNA

quantification in microbiome analysis. While microbial DNA was quantified through qPCR targeting the 16S rRNA gene, as there is no method for specifically quantifying CRISPR arrays, the stability of CRISPR analysis is likely affected by the copy number of CRISPR arrays. Future studies should focus on the development of quantification methods for CRISPR arrays, which may facilitate the determination of the quality of metagenomic CRISPR typing results, to better understand the CRISPR diversity in human skin environments and to control the quality of methodologies.

The EZ1 platform, an automated nucleic acid extraction platform, can be employed in forensic laboratories to improve the speed and reproducibility of DNA extraction from forensically relevant samples with forensic-grade quality. Therefore, utilization of the EZ1 protocol in microbiome analysis could be beneficial, as forensic analysts can share the same extracted DNA samples for both human and microbial DNA analysis, because most microbiome studies utilize DNA extraction protocols that are not forensically approved. Thus, this study proposes the utilization of the EZ1 protocol not only for forensic human DNA analysis but also for microbial DNA analysis for efficient sample consumption and convenience. The results of this chapter will be useful for future development of forensic microbiome analysis to achieve more informative human identification by complementing human STR typing from low-biomass samples, such as touched objects.

Summary

The human skin microbiome, a complex microbial community harbored by the human skin surface, has been proposed as a new forensic genetic marker for human identification to complement STR typing. *Streptococcus* CRISPR polymorphism of the skin microbiome was utilized to characterize individuality of the skin microbiome in chapter II; however, further validation is required for forensic purposes. Optimization of the DNA extraction protocol is critical for both forensic human DNA and microbiome analyses. Previous reports have demonstrated that DNA extraction protocols may alter the results of microbiome analysis, suggesting careful selection of DNA extraction protocols. In this study, protocols for DNA extraction from skin swab samples were evaluated by comparing two DNA extraction protocols, EZ1 DNA Investigator Kit and DNeasy PowerSoil Kit, in order to optimize the metagenomic CRISPR typing method. Quantification of 16S rRNA gene copy numbers and qualification of CRISPR diversities suggested that microbial DNA yield and composition of the CRISPRs were not altered by the two DNA extraction methods. Furthermore, the accuracy of the individual prediction models built using each sample set was comparable. Thus, these results support that the EZ1 platform, an automated DNA extraction protocol used in forensic laboratories, can also be used for metagenomic CRISPR typing intended for forensic human identification.

Table 9 Metadata for all samples newly collected in this chapter.

Sample ID	Participant ID	Date	Time Point	Gender	Age Range	Systemic usage in 6 months	antibiotics Steroids usage in months	6 Skin care products usage in 24 hours	Hand wash in 2 hours	Site-Symmetry	Site characteristic	Method
MCP002	P01	2019-07-02	1	Male	30s	No	Yes	Yes	Yes	Hp-L	Dry	EZ1
MCP004	P02	2019-07-02	1	Female	20s	Yes	No	Yes	Yes	Hp-L	Dry	EZ1
MCP006	P03	2019-07-03	1	Female	40s	No	No	No	Yes	Hp-L	Dry	EZ1
MCP007	P04	2019-08-07	2	Male	20s	No	No	No	Yes	Hp-L	Dry	EZ1
MCP009	P04	2019-08-07	2	Male	20s	No	No	No	Yes	Ac-L	Wet	EZ1
MCP011	P04	2019-08-07	2	Male	20s	No	No	No	Yes	Ra-L	Sebaceous	EZ1
MCP015	P01	2019-09-02	2	Male	30s	No	Yes	No	No	Hp-R	Dry	EZ1
MCP017	P01	2019-09-02	2	Male	30s	No	Yes	No	No	Ac-R	Wet	EZ1
MCP019	P01	2019-09-02	2	Male	30s	No	Yes	No	No	Ra-R	Sebaceous	EZ1
MCP021	P03	2019-09-02	2	Female	40s	No	No	Yes	Yes	Hp-R	Dry	EZ1
MCP022	P03	2019-09-02	2	Female	40s	No	No	Yes	Yes	Ac-L	Wet	EZ1
MCP025	P03	2019-09-02	2	Female	40s	No	No	Yes	Yes	Ra-R	Sebaceous	EZ1
MCP026	P02	2019-09-12	2	Female	20s	No	No	No	Yes	Hp-L	Dry	EZ1
MCP029	P02	2019-09-12	2	Female	20s	No	No	No	Yes	Ac-R	Wet	EZ1
MCP030	P02	2019-09-12	2	Female	20s	No	No	No	Yes	Ra-L	Sebaceous	EZ1
MCP032	P04	2019-12-02	3	Male	20s	No	No	No	No	Hp-L	Dry	PowerSoil
MCP033	P04	2019-12-02	3	Male	20s	No	No	No	No	Hp-R	Dry	EZ1
MCP034	P04	2019-12-02	3	Male	20s	No	No	No	No	Ac-L	Wet	PowerSoil
MCP035	P04	2019-12-02	3	Male	20s	No	No	No	No	Ac-R	Wet	EZ1
MCP036	P04	2019-12-02	3	Male	20s	No	No	No	No	Ra-L	Sebaceous	PowerSoil
MCP037	P04	2019-12-02	3	Male	20s	No	No	No	No	Ra-R	Sebaceous	EZ1
MCP038	P05	2019-12-12	1	Male	30s	Yes	No	No	Yes	Hp-L	Dry	EZ1
MCP040	P05	2019-12-12	1	Male	30s	Yes	No	No	Yes	Ac-L	Wet	EZ1
MCP042	P05	2019-12-12	1	Male	30s	Yes	No	No	Yes	Ra-L	Sebaceous	EZ1
MCP044	P03	2020-01-20	3	Female	40s	No	No	No	No	Hp-L	Dry	EZ1
MCP045	P03	2020-01-20	3	Female	40s	No	No	No	No	Hp-R	Dry	PowerSoil
MCP046	P03	2020-01-20	3	Female	40s	No	No	No	No	Ac-L	Wet	EZ1
MCP047	P03	2020-01-20	3	Female	40s	No	No	No	No	Ac-R	Wet	PowerSoil
MCP048	P03	2020-01-20	3	Female	40s	No	No	No	No	Ra-L	Sebaceous	PowerSoil
MCP049	P03	2020-01-20	3	Female	40s	No	No	No	No	Ra-R	Sebaceous	EZ1
MCP051	P01	2020-01-21	3	Male	30s	No	No	Yes	Yes	Hp-L	Dry	PowerSoil
MCP052	P01	2020-01-21	3	Male	30s	No	No	Yes	Yes	Hp-R	Dry	EZ1
MCP053	P01	2020-01-21	3	Male	30s	No	No	Yes	Yes	Ac-L	Wet	PowerSoil
MCP054	P01	2020-01-21	3	Male	30s	No	No	Yes	Yes	Ac-R	Wet	EZ1
MCP055	P01	2020-01-21	3	Male	30s	No	No	Yes	Yes	Ra-L	Sebaceous	PowerSoil
MCP056	P01	2020-01-21	3	Male	30s	No	No	Yes	Yes	Ra-R	Sebaceous	EZ1
MCP058	P05	2020-03-09	2	Male	30s	No	No	Yes	Yes	Hp-R	Dry	EZ1
MCP059	P05	2020-03-09	2	Male	30s	No	No	Yes	Yes	Ac-L	Wet	EZ1
MCP062	P05	2020-03-09	2	Male	30s	No	No	Yes	Yes	Ra-R	Sebaceous	EZ1
MCP063	EC	2020-07-03								EC	EC	EZ1

Table 10 Comparison of accuracies of individual prediction using different DNA extraction protocols.

CRISPR name	Sample Group	N	Accuracy (%)
CRISPR3	EZ1	25	80
	PS	30	80
	Overall	55	89
CRISPR1	EZ1	23	96
	PS	27	93
	Overall	50	96

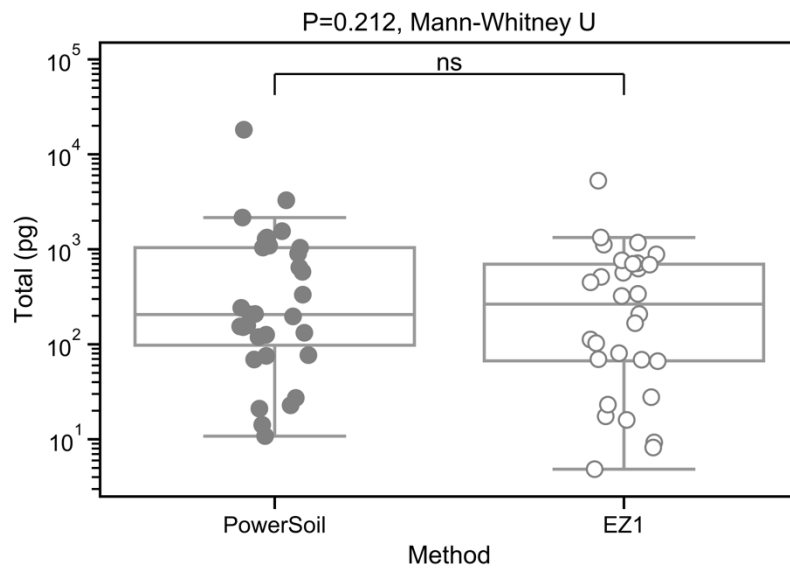


Figure 12 Comparison of DNA yields of different DNA extraction methods.

The total amount of microbial DNA extracted from skin swab samples using different DNA extraction methods was compared. The Mann-Whitney U test was used for pairwise comparison. ns, non-significant.

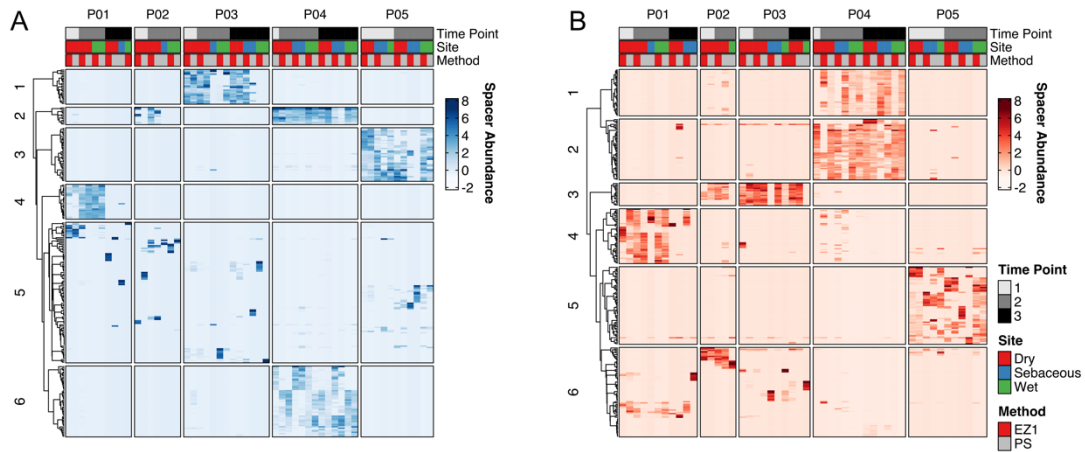


Figure 13 Visualization of personalized spacer diversity.

The spacer abundances of (A) CRISPR3 and (B) CRISPR1 were standardized as z-scores and then visualized as heatmaps. Columns represent samples split by individuals, while rows represent unique spacer sequence ASVs split into six blocks by spacer cluster based on dendrograms obtained using the Ward method. Top annotations denote sample type, including time point, skin site characteristics, and DNA extraction method. See also Table 6 and Table 9 for detailed information on the annotations.

The ASVs that were equal to or more than 200 times in total were used to generate the heatmaps.

CONCLUSION

While forensic personal identification is essential for criminal investigation, STR typing is difficult when the amount of human DNA recovered from the samples is too low. Microbiome-based measures may serve as an alternative approach for such low-biomass samples. In this thesis, metagenomic CRISPR typing which analyze diversity of *Streptococcus* CRISPRs was developed and assessed to achieve forensic personal identification.

In chapter I, 24 putative CRISPR arrays were identified by metagenomic reconstruction using publicly available datasets, as putative CRISPRs shared between the skin of different individuals. In addition, comparison of CRISPRs identified by metagenomic reconstruction revealed uniqueness of CRISPR arrays found in skin microbiome, suggesting that CRISPRs identified in chapter I would be useful marker for the characterization of personal microbiome

In chapter II, spacer sequences of three CRISPR loci were analyzed by amplicon sequencing. A total of 1223462 reads from three CRISPR arrays revealed that spacers in the skin microbiome were highly personalized, and conserved repeats were commonly shared between individuals. These individual specificities observed using CRISPR typing were confirmed by comparing the CRISPR diversity to microbiome diversity assessed using 16S rRNA amplicon sequencing. CRISPR typing achieved 95.2 % accuracy in personal classification, whereas 16S rRNA sequencing only achieved 52.6 %, suggesting that sequencing CRISPRs in the skin microbiome may be a more powerful approach for personal identification compared to 16S rRNA sequencing.

In chapter III, two DNA extraction protocols were compared, in order to optimize

the metagenomic CRISPR typing method. Quantification of 16S rRNA gene copy numbers and qualification of CRISPR diversities suggested that microbial DNA yield and composition of the CRISPRs were not altered by the two DNA extraction methods. Furthermore, the accuracy of the individual prediction models built using each sample set was comparable. Thus, the results provided in this chapter support that the automated DNA extraction protocol used in forensic laboratories can also be used for metagenomic CRISPR typing intended for personal identification.

Amplicon sequencing of CRISPRs identified by metagenomic reconstruction revealed highly personalized nature of streptococcal CRISPRs in skin microbiome. While this study is a proof-of-concept, CRISPR typing developed in the present thesis may be used as forensic tool to complement conventional DNA typing methods. Continuous effort on the optimization and validation of the CRISPR typing should be made to address the major issues remained for forensic application.

ACKNOWLEDGMENTS

First of all, I would like to express deep gratitude to Professor Hideaki Higashi (Division of Infection and Immunity, International Institute for Zoonosis Control, Hokkaido University) for giving me the opportunity to complete my thesis and for his support and guidance.

I would also like to express great appreciation to Professor Kazuhiko Ohashi (International Affairs Office, Faculty of Veterinary Medicine, Hokkaido University), Professor Yasuhiko Suzuki (Division of Bioresources, International Institute for Zoonosis Control, Hokkaido University) and Associate Professor Ryo Nakao (Laboratory of Parasitology, Faculty of Veterinary Medicine, Hokkaido University) for their critical review and valuable advice on the manuscript.

I would like to express deep appreciate to members of National Research Institute of Police Science, Japan, Mr. Satoshi Kubota, Dr. Ai Asahi, Dr. Tomoko Akutsu and Dr. Ken Watanabe for providing valuable advice and administration during conducting the project. I also greatly appreciate to Dr. Ryo Yokota, who inspired the original idea of this study and helped me in bioinformatic analysis. I also thank to Dr. Koji Fujii, Ms. Mai Kishimoto, and Dr. Manato Hirabayashi for their technical supports.

I would also like to acknowledge the funding from the Japan Society for the Promotion of Science (JSPS). The work of this study was funded in part by a Grant-in-Aid for Research Activity Start-up (19K24245) and a Grant-in-Aid for Early-Career Scientists (19K20408 and 20K18991).

Finally, I would like to thank my family for supporting me in a wide variety of ways and providing me the motivation I needed to continue the works.

和文要旨

近年の DNA 配列解析技術の革新に伴ってマイクロバイーム研究が急速に進展し、ヒトの体表を含めた環境中には、環境毎にその系統や組成が異なった微生物の集団が生存していることが明らかにされてきている。法科学領域においてはマイクロバイームの構造が環境によって異なることを応用して、ヒト皮膚マイクロバイーム解析による個人識別法の開発を目指す研究が各国で進められている。近年のマイクロバイーム研究は大きく二つの手法によって実施されており、一つがアンプリコンシーケンス解析、他方がショットガンメタゲノム解析である。従来用いられることが多かったアンプリコンシーケンス解析は、PCR により 16S rRNA 配列などのマーカー遺伝子を増幅し、網羅的に解読する手法である。しかしながら 16S rRNA 配列は近縁種では類似しており、しばしば属単位までの系統情報しか得られないことがある。そのため、個人識別においては解像度が低いと言われてきた。他方、ショットガンメタゲノム解析は試料中の DNA を断片化し、これを網羅的に解読する手法で、高い解像度でマイクロバイームの構造を把握することができる。しかし、解析にかかるコストが高く、PCR による増幅を経ないために多量の DNA が必要となるといった難点を抱える。このような背景から、ヒト皮膚マイクロバイーム解析による個人識別法の実現には、16S rRNA 配列よりも高い解像度で微生物の群集構造を捉えるマーカー遺伝子を利用した、新規アンプリコンシーケンス解析法を開発する必要があると考えられる。

CRISPR は原核生物の持つリピート配列で、Cas タンパク質と共に原核生物における「獲得免疫系」を担う。CRISPR はその機能を果たす上で、細胞内へ侵入してきたファージを中心とする可動性遺伝因子の DNA 配列を、リピート配列中にスペーサーと呼ばれる配列として取り込む性質を持つ。このような性質から CRISPR 配列の多様性は環境中のウイルス叢の構造を反映している。ヒトの皮膚

におけるウイルス叢は細菌叢に比べて個人間の多様性が大きいことが知られており、本研究では CRISPR をマーカーとする新規アンプリコンシーケンス解析法を開発し、CRISPR の多様性に基づいた個人識別の法科学領域における有用性を検討することを目的とした。

第 I 章では、公開データベースに登録されたデータセットを利用して、皮膚マイクロバイオーム中に存在する CRISPR を再構築し、複数人の間で共有される CRISPR のリピート配列を同定した。また、再構築された CRISPR について、その配列をリファレンスとなる CRISPR と比較したところ、再構築された CRISPR のスペーサー配列はいずれもリファレンス CRISPR には存在しない独自の配列であり、マーカーとしての有用性が示唆された。

第 II 章では、3 種類の候補 CRISPR について、スペーサー配列をアンプリコンシーケンスにより解析した。性能評価のために 16S rRNA 配列と比較したところ、スペーサー配列の多様性は 16S rRNA 配列の多様性より個人間の多様性が大きく、高い正答率で個人を識別可能であった。このことは CRISPR の配列解析が 16S rRNA の配列解析に比べ、個人識別用途に優れる可能性を示唆するものである。

第 III 章では、解析系の DNA 抽出法を最適化するために、第 II 章で使用した細菌叢解析に使用される DNA 抽出法と、自動抽出装置を用いた法科学領域で使用される DNA 抽出法の、2 種類の DNA 抽出法を比較した。16S rRNA 配列のコピー数の定量と CRISPR の多様性評価の結果、2 種類の DNA 抽出法は細菌 DNA の抽出効率について同等であり、CRISPR の多様性は DNA 抽出法によらず安定していると考えられた。また、試料の識別精度も同等であり、法科学領域で使用される自動抽出装置によって抽出された DNA が CRISPR の多様性評価にも使用可能であることが示された。

メタゲノムデータセットから同定された CRISPR のアンプリコンシーケンス

解析により、皮膚上の連鎖球菌に由来する CRISPR の多様性が個人特異的であることが明らかになった。本研究は概念実証ではあるが、開発した CRISPR 解析法は従来の DNA 型検査法を補完する手法として、法科学領域において応用可能であると考えられる。今後もさらなる解析系の最適化と法科学的評価進めることで、実用化に向けた課題の解決を目指し、さらなる解析系の最適化と法科学的評価が必要である。

REFERENCE

1. Marchesi JR, Ravel J. 2015. The vocabulary of microbiome research: a proposal. *Microbiome* 3:31.
2. Human Microbiome Project, Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207-214.
3. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI. 2008. Evolution of mammals and their gut microbes. *Science* 320:1647-1651.
4. Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, Siegel J, Caporaso JG. 2016. Geography and Location Are the Primary Drivers of Office Microbiome Composition. *mSystems* 11:e00022-16.
5. Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasen AS, Pedersen BK, Al-Soud WA, Sorensen SJ, Hansen LH, Jakobsen M. 2010. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* 5:e9085.
6. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, Nomicos E, Polley EC, Komarow HD, Program NCS, Murray PR, Turner ML, Segre JA. 2012. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res* 22:850-859.
7. Urushiyama D, Suda W, Ohnishi E, Araki R, Kiyoshima C, Kurakazu M, Sanui A, Yotsumoto F, Murata M, Nabeshima K, Yasunaga S, Saito S, Nomiyama M, Hattori M, Miyamoto S, Hata K. 2017. Microbiome profile of the amniotic fluid as a predictive biomarker of perinatal outcome. *Sci Rep* 7:12171.

8. Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59:143-169.
9. Byrd AL, Belkaid Y, Segre JA. 2018. The human skin microbiome. *Nat Rev Microbiol* 16:143-155.
10. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolk T, McCall LI, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. *Nat Rev Microbiol* 16:410-422.
11. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A* 107:6477-6481.
12. 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.
13. Oh J, Byrd AL, Park M, Program NCS, Kong HH, Segre JA. 2016. Temporal Stability of the Human Skin Microbiome. *Cell* 165:854-866.
14. Fierer N, Hamady M, Lauber CL, Knight R. 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci U S A* 105:17994-17999.
15. Schmedes SE, Woerner AE, Budowle B. 2017. Forensic human identification using skin microbiomes. *Appl Environ Microbiol* 83:e01672-17.
16. 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.
17. 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.
 18. Daly DJ, Murphy C, McDermott SD. 2012. The transfer of touch DNA from hands to glass, fabric and wood. *Forensic Sci Int Genet* 6:41-46.
 19. National Police Agency, Japan. 2021. Keisatsu hakusho (White Paper on Police 2021).112.
 20. Costello EEK, Lauber CCL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. *Science* 326:1694-1697.
 21. Quaak FCA, van Duijn T, Hoogenboom J, Kloosterman AD, Kuiper I. 2018. Human-associated microbial populations as evidence in forensic casework. *Forensic Sci Int Genet* 36:176-185.
 22. Benschop CCG, Quaak FCA, Boon ME, Sijen T, Kuiper I. 2012. Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? *Int J Legal Med* 126:303-310.
 23. Nishi E, Tashiro Y, Sakai K. 2015. Discrimination among individuals using terminal restriction fragment length polymorphism profiling of bacteria derived from forensic evidence. *Int J Legal Med* 129:425-433.
 24. Wilkins D, Leung MHY, Lee PKH. 2017. Microbiota fingerprints lose individually identifying features over time. *Microbiome* 5:1.

25. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannon BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci U S A* 112:E2930-E2938.
26. Ghebremedhin B, Layer F, König W, König B. 2008. Genetic classification and distinguishing of *Staphylococcus* species based on different partial gap, 16S rRNA, hsp60, rpoB, sodA, and tuf gene sequences. *J Clin Microbiol* 46:1019-1025.
27. 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.
28. Bhaya D, Davison M, Barrangou R. 2011. CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. *Annu Rev Genet* 45:273-297.
29. Barrangou R, Marraffini LA. 2014. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol Cell* 54:234-244.
30. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero Da, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709-1712.
31. Abeles SR, Robles-Sikisaka R, Ly M, Lum AG, Salzman J, Boehm TK, Pride DT. 2014. Human oral viruses are personal, persistent and gender-consistent. *ISME J* 8:1753-1767.
32. Naidu M, Robles-Sikisaka R, Abeles SR, Boehm TK, Pride DT. 2014. Characterization of bacteriophage communities and CRISPR profiles from dental

- plaque. BMC Microbiol 14:175.
33. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35:907-914.
 34. Hoe N, Nakashima K, Grigsby D, Pan X, Dou SJ, Naidich S, Garcia M, Kahn E, Bergmire-Sweat D, Musser JM. 1999. Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. Emerg Infect Dis 5:254-263.
 35. Schouls LM, Reulen S, Duim B, Wagenaar JA, Willems RJ, Dingle KE, Colles FM, Van Embden JD. 2003. Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. J Clin Microbiol 41:15-26.
 36. Pourcel C, Salvignol G, Vergnaud G. 2005. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151:653-663.
 37. Mokrousov I, Narvskaya O, Limeschenko E, Vyazovaya A. 2005. Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. J Clin Microbiol 43:1662-1668.
 38. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. J Bacteriol 190:1401-1412.
 39. Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de Romans S, Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-

- Jeanjean S, Achtman M, Brisse S, Sola C, Weill FX. 2012. CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One* 7:e36995.
40. Lier C, Baticle E, Horvath P, Haguenoer E, Valentin AS, Glaser P, Mereghetti L, Lanotte P. 2015. Analysis of the type II-A CRISPR-Cas system of *Streptococcus agalactiae* reveals distinctive features according to genetic lineages. *Front Genet* 6:214.
 41. Beauruelle C, Pastuszka A, Horvath P, Perrotin F, Mereghetti L, Lanotte P. 2017. CRISPR: A useful genetic feature to follow vaginal carriage of group B *Streptococcus*. *Front Microbiol* 8:1981.
 42. Hu T, Cui Y, Qu X. 2020. Characterization and comparison of CRISPR Loci in *Streptococcus thermophilus*. *Arch Microbiol* 202:695-710.
 43. Shariat N, Dudley EG. 2014. CRISPRs: molecular signatures used for pathogen subtyping. *Appl Environ Microbiol* 80:430-439.
 44. Sun CL, Thomas BC, Barrangou R, Banfield JF. 2016. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J* 10:858-870.
 45. Rho M, Wu Y-W, Tang H, Doak TG, Ye Y. 2012. Diverse CRISPRs Evolving in Human Microbiomes. *PLoS Genetics* 8:e1002441.
 46. Gogleva AA, Gelfand MS, Artamonova II. 2014. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics* 15:202.
 47. Paez-Espino D, Eloie-Fadrosch EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's

- virome. *Nature* 536:425-430.
48. Pride DT, Sun CL, Salzman J, Rao N, Loomer P, Armitage GC, Banfield JF, Relman DA. 2011. Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res* 21:126-136.
 49. Robles-Sikisaka R, Ly M, Boehm T, Naidu M, Salzman J, Pride DT. 2013. Association between living environment and human oral viral ecology. *ISME J* 7:1710-1724.
 50. Robles-Sikisaka R, Naidu M, Ly M, Salzman J, Abeles SR, Boehm TK, Pride DT. 2014. Conservation of streptococcal CRISPRs on human skin and saliva. *BMC Microbiol* 14:146.
 51. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.
 52. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674-1676.
 53. Edgar RC. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18.
 54. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodriguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S,

- Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37:852-857.
55. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584.
 56. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
 57. Godde JS, Bickerton A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. *J Mol Evol* 62:718-729.
 58. Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 9:90-95.
 59. Pourcel C, Touchon M, Villeriot N, Vernadet JP, Couvin D, Toffano-Nioche C, Vergnaud G. 2020. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res* 48:D535-D544.
 60. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Neron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 46:W246-W251.
 61. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, Moineau S, Mojica FJM, Scott D,

- Shah SA, Siksnyš V, Terns MP, Venclovas C, White MF, Yakunin AF, Yan W, Zhang F, Garrett RA, Backofen R, van der Oost J, Barrangou R, Koonin EV. 2020. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 18:67-83.
62. Nethery MA, Barrangou R. 2019. CRISPR Visualizer: rapid identification and visualization of CRISPR loci via an automated high-throughput processing pipeline. *RNA Biol* 16:577-584.
 63. Hao M, Cui Y, Qu X. 2018. Analysis of CRISPR-Cas System in *Streptococcus thermophilus* and Its Application. *Front Microbiol* 9:257.
 64. Liu J, Yan R, Zhong Q, Ngo S, Bangayan NJ, Nguyen L, Lui T, Liu M, Erfe MC, Craft N, Tomida S, Li H. 2015. The diversity and host interactions of *Propionibacterium acnes* bacteriophages on human skin. *ISME J* 9:2078-2093.
 65. Fitz-Gibbon S, Tomida S, Chiu B-H, 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* 133:2152-2160.
 66. Brüggemann H, Lomholt HB, Tettelin H, Kilian M. 2012. CRISPR/cas loci of type II *Propionibacterium acnes* confer immunity against acquisition of mobile elements present in type I *P. acnes*. *PLoS One* 7:e34171.
 67. Acuna-Amador L, Primot A, Cadieu E, Roulet A, Barloy-Hubler F. 2018. Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of *Porphyromonas gingivalis* reference strains. *BMC Genomics* 19:54.
 68. Pride DT, Salzman J, Relman DA. 2012. Comparisons of clustered regularly

- interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses. *Environ Microbiol* 14:2564-2576.
69. Lum AG, Ly M, Santiago-Rodriguez TM, Naidu M, Boehm TK, Pride DT. 2015. Global transcription of CRISPR loci in the human oral cavity. *BMC Genomics* 16:401.
 70. Castelino M, Eyre S, Moat J, Fox G, Martin P, Ho P, Upton M, Barton A. 2017. Optimisation of methods for bacterial skin microbiome investigation: Primer selection and comparison of the 454 versus MiSeq platform. *BMC Microbiol* 17:23.
 71. Zheng W, Tsompana M, Ruscitto A, Sharma A, Genco R, Sun Y, Buck MJ. 2015. An accurate and efficient experimental approach for characterization of the complex oral microbiota. *Microbiome* 3:48.
 72. Kong HH, Andersson B, Clavel T, Common JE, Jackson SA, Olson ND, Segre JA, Traidl-Hoffmann C. 2017. Performing Skin Microbiome Research: A Method to the Madness. *J Invest Dermatol* 137:561-568.
 73. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581-583.
 74. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27.
 75. Kruskal WH, Wallis WA. 1952. Use of Ranks in One-Criterion Variance Analysis. *J Am Statist Assoc* 47:583-621.

76. Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O' Hara R, Simpson G, Solymos P, Stevens M, Szoecs E, Wagner H. 2019. Vegan: Community Ecology Package, version 2.5-4, <https://CRAN.R-project.org/package=vegan>.
77. Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46:389–422.
78. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12:2825-2830.
79. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590-596.
80. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12:87.
81. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11:2639-2643.
82. Gotelli NJ, Chao A. 2013. Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data, p 195-211. *In* Levin S (ed), *Encyclopedia of Biodiversity* doi:10.1016/B978-0-12-384719-5.00424-X. Academic Press.
83. Cottier F, Srinivasan KG, Yurieva M, Liao W, Poidinger M, Zolezzi F, Pavelka N.

2018. Advantages of meta-total RNA sequencing (MeTRS) over shotgun metagenomics and amplicon-based sequencing in the profiling of complex microbial communities. *NPJ Biofilms Microbiomes* 4:2.
84. McGinn J, Marraffini LA. 2019. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat Rev Microbiol* 17:7-12.
 85. Rascovan N, Duraisamy R, Desnues C. 2016. Metagenomics and the Human Virome in Asymptomatic Individuals. *Annu Rev Microbiol* 70:125-141.
 86. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. *FEMS Microbiol Rev* 35:343-359.
 87. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, Pei Z, Blaser MJ, Aliferis CF, Alekseyenko AV. 2013. A comprehensive evaluation of multiclassification methods for microbiomic data. *Microbiome* 1:11.
 88. 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.
 89. Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF. 2015. CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio* 6:e00262-15.
 90. Kutmutia SK, Drautz-Moses DI, Uchida A, Purbojati RW, Wong A, Kushwaha KK, Putra A, Premkrishnan BNV, Heinle CE, Vettath VK, Junqueira ACM, Schuster SC. 2019. Complete Genome Sequence of *Micrococcus luteus* Strain SGAir0127, Isolated from Indoor Air Samples from Singapore. *Microbiol Resour Announc* 8:e00646-19.
 91. Tomida S, Nguyen L, Chiu BH, Liu J, Sodergren E, Weinstock GM, Li H. 2013.

- Pan-Genome and Comparative Genome Analyses of *Propionibacterium acnes* Reveal Its Genomic Diversity in the Healthy and Diseased Human Skin Microbiome. *mBio* 4:e00003-13.
92. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 95:3140-3145.
 93. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. 2006. Proposal for standardization of optimized *Mycobacterium* interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J Clin Microbiol* 44:4498-4510.
 94. Zolfo M, Tett A, Jousson O, Donati C, Segata N. 2017. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res* 45:e7.
 95. Shevchenko SG, Radey M, Tchesnokova V, Kisiela D, Sokurenko EV. 2019. *Escherichia coli* Clonobiome: Assessing the Strain Diversity in Feces and Urine by Deep Amplicon Sequencing. *Appl Environ Microbiol* 85:e01866-19.
 96. Meadow JF, Altrichter AE, Green JL. 2014. Mobile phones carry the personal microbiome of their owners. *PeerJ* 2:e447.
 97. Hidalgo-Cantabrana C, Sanozky-Dawes R, Barrangou R. 2018. Insights into the human virome using CRISPR spacers from microbiomes. *Viruses* 10:479.
 98. Nadkarni MA, Martin FE, Jacques NA, Hunter N. 2002. Determination of

- bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* 148:257-266.
99. Li R. 2015. Autosomal Short Tandem Repeat Profiling, p 369-406, *Forensic Biology*, 2 ed. CRC Press.
 100. 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.
 101. Nishi E, Watanabe K, Tashiro Y, Sakai K. 2017. Terminal restriction fragment length polymorphism profiling of bacterial flora derived from single human hair shafts can discriminate individuals. *Leg Med* 25:75-82.
 102. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. 2012. Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. *PLoS One* 7:e33865.
 103. Hornung BVH, Zwittink RD, Kuijper EJ. 2019. Issues and current standards of controls in microbiome research. *FEMS Microbiol Ecol* 95:fiz045.
 104. Bjerre RD, Hugerth LW, Boulund F, Seifert M, Johansen JD, Engstrand L. 2019. Effects of sampling strategy and DNA extraction on human skin microbiome investigations. *Sci Rep* 9:17287.
 105. Davis CP, King JL, Budowle B, Eisenberg AJ, Turnbough MA. 2012. Extraction platform evaluations: a comparison of AutoMate Express, EZ1(R) Advanced XL, and Maxwell(R) 16 Bench-top DNA extraction systems. *Leg Med* 14:36-39.
 106. Fujii K, Inokuchi S, Kitayama T, Nakahara H, Mizuno N, Sekiguchi K. 2013. A comparison of DNA extraction using AutoMate Express and EZ1 advanced XL from liquid blood, bloodstains, and semen stains. *J Forensic Sci* 58:981-988.

107. Akutsu T, Kitayama T, Watanabe K, Sakurada K. 2015. Comparison of automated and manual purification of total RNA for mRNA-based identification of body fluids. *Forensic Sci Int Genet* 14:11-17.
108. Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32:2847-2849.
109. Fiedorova K, Radvansky M, Nemcova E, Grombirikova H, Bosak J, Cernochova M, Lexa M, Smajs D, Freiburger T. 2019. The Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community Recovery. *Front Microbiol* 10:821.
110. Lim MY, Song EJ, Kim SH, Lee J, Nam YD. 2018. Comparison of DNA extraction methods for human gut microbial community profiling. *Syst Appl Microbiol* 41:151-157.
111. Erb-Downward JR, Falkowski NR, D'Souza JC, McCloskey LM, McDonald RA, Brown CA, Shedden K, Dickson RP, Freeman CM, Stringer KA, Foxman B, Huffnagle GB, Curtis JL, Adar SD. 2020. Critical Relevance of Stochastic Effects on Low-Bacterial-Biomass 16S rRNA Gene Analysis. *mBio* 11:e00258-20.