Forensic Science International: Genetics 57 (2022) 102662

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

The application of the skin virome for human identification

Ema H. Graham [a], Jennifer L. Clarke [b], Samodha C. Fernando [c], Joshua R. Herr [d], Michael S. Adamowicz [e,*]

[a] Program in Complex Biosystems, University of Nebraska, 68583 Lincoln, NE, USA
[b] Department of Statistics, University of Nebraska, Lincoln, NE 68588, USA
[c] Department of Animal Science, University of Nebraska, Lincoln, NE 68583, USA
[d] Department of Plant Pathology and Center for Plant Science Innovation, University of Nebraska, Lincoln, NE 68503, USA
[e] College of Agricultural Sciences & Natural Resources, University of Nebraska, Lincoln, NE 68583, USA

ARTICLE INFO

Keywords:
Virome
Metagenome
Human identification
Viral biomarker
Human skin

ABSTRACT

The use of skin virome offers a unique approach for human identification purposes in instances where a viable and statistically relevant human DNA profile is unavailable. The skin virome may act as an alternative DNA profile and/or an additional form of probative genetic material. To date, no study has attempted to investigate the human virome over a time series across various physical locations of the body to identify its diagnostic potential as a tool for human identification. For this study, we set out to evaluate the stability, diversity, and individualization of the human skin virome. An additional goal was to identify putative viral signatures that can be used in conjunction with traditional forensic STR loci. In order to accomplish this, human viral metagenomes were collected and sequenced from 42 individuals at three anatomical locations (left hand, right hand, and scalp) across multiple collection periods over a 6-month window of time. Assembly dependent and independent bioinformatic approaches, along with a database centered assessment of viral identification, resulted in three sets of stable putative viral markers. In total, with the three sets combined, we identified 59 viral biomarker regions, consisting of viral species and uncharacterized viral genome assemblies, that were stable over the sampling period. Additionally, we found the abundance profiles of these 59 viral biomarkers, based on presence or absence, to be significantly different across subjects ($P < 0.001$). Here we demonstrate that not only is the human virome applicable to be used for human identification, but we have identified many viral signatures that can putatively be used for forensic applications, thus providing a foundation to the novel field of forensic virology.

1. Introduction

The use of microbial signatures for human identification constitutes a new form of information that can be used in forensic applications. The wide variety of data types, sample locations, environmental factors, analysis methods, and the diversity of microorganisms make this emerging area of forensic biology increasingly relevant [1]. Although the experimental study of microbiome analysis in relation to casework is relatively new, the use of bacteria in forensically relevant cases dates back to the late 1800s [2]. In one modern example, researchers have examined commercial honey products for the presence of *Clostridium botulinum* as a vector for cases of induced poisoning from botulism [3]. More recently, *Neisseria gonorrhoeae* infections in children have been considered as evidence in sexual assault cases [4]. As such, the application of microbes in forensic science is not a new concept.

The potential of utilizing the human microbiome in forensic science has previously been described by Hampton-Marcell et al. [5] who performed experiments to identify associations between human skin bacterial composition and the physical environment. That study [5] concluded that using the bacterial 16S ribosomal RNA (rRNA) marker region provided weak predictions of human identity compared to more traditional methods of forensic human DNA analysis. However, they also recognized that there could be better methods of analysis and that the field of study is currently in its infancy [5]. In a recent study addressing the human skin bacterial microbiome for the purpose of human identification, researchers identified a set of bacterial taxa that were able to classify individuals beyond a one-year period with up to 85% accuracy [6]. Additional work conducted by Schmedes et al. [7] used publicly available shotgun metagenomic data collected from 12 human individuals spanning multiple skin site locations. Their findings identified

* Corresponding author.
E-mail address: madamowicz2@unl.edu (M.S. Adamowicz).

https://doi.org/10.1016/j.fsigen.2022.102662
Received 13 September 2021; Received in revised form 14 January 2022; Accepted 16 January 2022
Available online 19 January 2022

multiple bacterial clade specific and single nucleotide variant markers from the bacterium *Propionibacterium acnes* that identified the 12 study participants with up to 100% accuracy [7].

While there has been ample work using bacteria as targets for forensic identification, there have been very few studies addressing the application of viruses and the human virome for forensic purposes. A potential method for tracing unidentified human cadavers using the JC virus was published by Ikegaya et al. [8] where the authors suggested that the presence of this virus in humans may aid in determining the geographical location of origin of the decedent. Wilson et al. [9] described specific considerations and presented seven discrete steps for analysis of viral samples in a forensic case. They also described a broad range of topics relevant to working with viral biomarkers, including considering viruses as weapons and the intentional/unintentional transmission of virus-mediated diseases in criminal cases. Additionally, the hidSkinPlex, a panel of microbial clade-specific markers which were developed for the assessment of the skin microbiome for human identification, utilizes the *Propionibacterium* phage P101A as a marker [10]. The development and validation study of this panel evaluated a small set of viral markers specific to *Propionibacterium* phage isolates; however, this study only utilized a single bacteriophage virus and did not address other viral types. Authors in this study acknowledged there is potential for viral markers to be used for human identification, however, they point out the need for further studies and investigation into phage and viral marker development utilizing viral targeted extraction and improved amplification techniques [10].

The human skin virome offers a unique advantage over just using bacterial based microbiome markers. The community composition of human skin bacteria is affected by a number of confounding factors, such as antibiotic intervention or the use of antibacterial soap [11]. For instances like these, alternative microbial markers must be used as they provide an alternative form of data to that of pre-established bacterial markers. Targeting viral populations that are part of the core skin virome, such as eukaryotic infecting viral populations not affected by antibacterial agents, not only offers additional biomarkers to those already established but potentially offers increased stability and detection even in the presence of outside environmental contributory factors.

To date, no study has attempted to exclusively investigate the human virome over a time series across various physical locations of the body to identify its potential as a tool for human identification separate from that of the human bacterial microbiome. One reason for the limited number of virome studies has been the lack of bioinformatic and molecular tools for human virome investigation. Despite this, the development of high-throughput sequencing methods and the resulting decrease in sequencing costs have led to studies investigating the viromes of humans, primarily focusing on describing viral diversity in the human gut [12–17]. These studies have demonstrated that human gut viromes tend to be "highly individual and temporally stable" [16], two key features that are highly desirable for a forensic biomarker.

In order to address the human skin virome in a forensic context, we investigated the temporal human skin virome stability on three body locations (left hand, right hand, and scalp) in 42 study participants using five longitudinal samples taken across a 6-month time period. The goal of this study was to address two hypotheses. First, we expect that the human skin virome consists of both stable and variable viral taxa but are unsure which viral taxa may be in each category. Second, we hypothesize that there will be viral taxa that may be diagnostic for human individualization and that we may attribute sampling location or lifestyle traits as a proxy to explain diagnostic differences in the human skin virome. Our overall goal in this exploratory study was to identify unique but stable viral sub-populations within the human skin virome and to assess this virome diversity for potential forensic human identification.

2. Materials and methods

2.1. Sample collection

Samples from the skin virome of the participants were collected using a tandem dry and wet swab technique using nylon flocked swabs (4NG Floq Swabs, Copan, Brescia, Italy), as previously described [18,19]. The wet swab was moistened with sterile 1x phosphate buffered saline (PBS) and acted as the leading swab. The wet swab was immediately followed by the dry swab to collect loosened sample material. Post swabbing, all collected swabs were stored in pre-sterilized Eppendorf safe lock 2 ml tubes (Eppendorf, Hamburg, Germany). Virome samples were collected from 42 adult individuals with ages spanning 19–70+ years, who were not on antibiotic drug regimens during the duration of the sampling. Originally 43 participants were a part of this study, however, one subject (P33) was unable to complete the study due to the required use of antibiotics during the duration of sampling time points and was therefore not used for the analysis of this study. Samples were collected across a longitudinal 6-month period to represent the initial sampling (day 0), and 2-weeks, 1-month, 3-months, and 6-months intervals from the initial sampling date. At each collection, virome swabs were collected from three skin locations – left hand, right hand, and scalp. At each sampling date the participants filled out a questionnaire to gather information on travel, skin care, lifestyle, and other information that could help identify factors affecting the skin viral microbiome. A representative questionnaire used in the study is included as Supplementary Fig. 1. During each sampling, negative control swabs were collected to evaluate any contamination that may have resulted from laboratory factors, including the batch of PBS used and deoxyribonucleic acid (DNA) extraction date, as well as subsequent sequencing run effects. Forensic samples would most likely not consist of highly degradable ribonucleic acid (RNA), so this initial study focused on DNA viruses, and as such, the collected swabs were stored at −20 °C until used for viral enrichment and sequencing.

2.2. Viral enrichment and purification

Swabs containing skin virome samples were saturated with 200 μl of 0.02 μm filtered sterile 1X PBS and were placed in a 2 ml tube containing a CW Spin Basket (Promega, Madison, WI, USA). Swabs were centrifuged at 16,000xg for 10 min to elute viral particles from the swab into the PBS solution. The filtrate containing the viral particles was further filtered using a 0.22 μm filter to remove cellular and bacterial contaminants. Enriched for viral particles, the resulting filtrate was used for viral DNA extraction using the QiAmp Ultra-Sensitive Virus Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol.

The resulting viral DNA was subjected to whole genome amplification (WGA) using multiple displacement amplification (MDA) implemented with the TruePrime WGA Kit (Syngen Biotechnology, Inc, Taipei City, Taiwan) following the standard product protocol. Following WGA, the samples were quantified using the DeNovix dsDNA High Sensitivity Kit using the DeNovix DS-11 Spectrophotometer/Fluorometer (DeNovix, Inc, Wilmington, DE, USA).

2.3. Viral metagenome library preparation and sequencing

One hundred nanograms of the amplified DNA was used for library preparation. The DNA was sheared using sonication to a mean length distribution of 600 bp. Sonication was performed using a Bioruptor (Diagenode, Denville, NJ, USA) with three cycles of 30 s on and 90 s off as per manufacturer instructions. The resulting sheared DNA was used for library preparation using the NEBNext Ultra II Library preparation kit (New England Biolabs, Ipswich, MA, USA) according to the manufacturer's protocol. During the kit process, recommended conditions for size selection of adapter-ligated DNA approximate insert size of 500–700 bp was used, followed by PCR enrichment of adapter ligated

DNA with five cycles of denaturation and annealing/extension. Final library preps were evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc, Santa Clara, CA, USA) with high sensitivity chips to identify sample base pair distribution and sample concentration. Additionally, libraries were quantified using the DeNovix dsDNA High Sensitivity Kit (DeNovix, Inc, Wilmington, DE, USA). The libraries were then sequenced using the 150 bp paired-end sequencing strategy on the Illumina HiSeq 2500 platform (Illumina, Inc, San Diego, CA, USA). All raw sequencing data has been deposited in the NCBI Short Read Archive (SRA) under the accession code PRJNA754140.

2.4. Assembly of metagenomic sequencing reads

The resulting sequencing data in Fastq format was evaluated for quality using FastQC v.0.11.9 [20] and were trimmed and filtered using the adaptive trimming tool Sickle v.1.3.3 [21] to remove low quality reads using a quality filter threshold of Q30 and a length threshold of 75 bp. Reads resulting from the PhiX spike-in were removed from trimmed reads using the BBDuk command with standard operational flags from the BBMap suite of tools [22]. Following quality filtering, bacterial contamination was assessed by mapping trimmed reads to the Silva 16S ribosomal database v.138.1 using BBMap flags and parameters described for high precision mapping of contamination detection as suggested in the BBMap Guide [22]. Additionally, all sample reads were mapped to the human genome (hg19) using BBMap with standard operational flags [22] for high precision mapping with low sensitivity in order to lower the risk of false positive mapping. All human genome mapped reads were removed to mitigate human host sequence contamination. Metagenome assemblies were performed using MEGAHIT v.1.2.8 [23]. Assemblies were performed using two approaches, 1) assembly within each sample, and 2) a master meta-assembly using all reads. Assembly quality was assessed using QUAST v.5.0.2 [24]. In addition, a meta-assembly using all negative control reads was performed to identify potential contaminants in the dataset that may arise from reagents. The virome assemblies generated were mapped to the negative control contigs greater than 1000 bp using BWA-mem [25] to remove reads that may have resulted from contamination. Subsequent contigs greater than 1000 bp were then utilized for downstream analysis of viral identification, diversity analysis, and assessment of stability of the virome.

2.5. Viral identification and taxonomic classification

Putative viral contigs – designated as contigs containing annotated viral genes – were identified using the tool CheckV v.0.7.0 [26]. Contigs first identified as viral via CheckV were additionally subjected to other classification schemes using various viral annotation and classification tools as described below. Viral contigs were classified using both nucleotide-based classification tools, such as Kraken2 v.2.0.8-beta [27], Demovir [28], and Blastn (with a >10% query coverage cut-off) [29], and also using a classification tool based on protein coding sequences, Kaiju v.1.7 [30]. We used the least common ancestor of the consensus hits from these tools for a small amount of the viral contigs in cases where the classification results had similar e-values or percent confidence (based on each respective bioinformatic tool) but different taxonomic results at the viral genus or species. This was done to reduce misclassification of viral contigs resulting from different algorithms and reference databases used across these bioinformatic tools. With all of the classification tools, the resulting classification used was based on those hits having the lowest e-value and/or highest percent confidence.

2.6. Mapping of raw reads to assemblies and viral contigs

The raw sample data, in the form of forward and reverse paired-end reads, were mapped to the metagenome assembly consisting of contigs from the meta-assembly. Read mapping was performed using Bowtie 2 v.2.3.5 [31]. The total sequence counts for all the contigs, the respective contig lengths, and the mapped sequence counts were used to normalize each sample measured as the RPKM (reads per kilobase per million) value. In the rare instance when a read had equivalent "best hits" as determined by Bowtie 2 in the reference contigs, the first "best" contig hit was retained as per the default Bowtie 2 settings. Subsequently, SAMtools v.1.9 [32] was used to generate a read abundance table associated with each viral contig for each of the respective samples. This merged table of all the sample read abundances, the equivalent of a traditional taxonomic unit count table, was further analyzed using "R" v.3.6.3 [33]. Unique contigs were used to identify viral taxonomic diversity and temporal changes. Diversity analysis using the read abundance data was performed using the Phyloseq R package as further described in Section 2.9 [34].

2.7. Evaluation of subject viral diversity abundance across skin sites

Of the identified viral contigs, abundance of all taxonomically assigned viral families were summed across all time points by anatomical sampling location within a subject. The relative abundance of the top ten most abundant viral families by location were identified and evaluated using the Phyloseq package. To compare abundance profiles within subjects across the sampled skin sites (i.e., left and right hand, and scalp) and between subjects, a Bray Curtis dissimilarity matrix was produced with the resulting distance of 0 being all classes being the same and 1 being all classes being disjointed. The subsequent Bray Curtis distances were averaged between each of the three subject's sampled anatomical location abundance profiles and all other abundance profiles (intersubject dissimilarity). Additionally, all distances between the three subjects' anatomical location abundance profiles were averaged (intrasubject dissimilarity). A series of pairwise Wilcoxon rank sum tests were performed between the intrasubject dissimilarity distances and the intersubject dissimilarity distances to assess differences in viral family community abundance across location sites within a subject and across subjects.

2.8. Assessment of virome stability over time

Viral contig stability was assessed on the basis of presence and absence of each viral contig over time across each body site sampled (left hand, right hand, or scalp) within each subject. Additionally, viral contig stability was assessed at the taxonomic level of viral family, genus, and species. Contigs present in four out of the five time points from a specific location within an individual were considered to be a stable viral contig and were identified as a potential marker for human identification. Our value of four samples as a criterion for stability was based on previous research [35] and, while arbitrary, this designation allowed us to account for random aberrations in sequencing data from our sequence provider, individuals who failed to show up for their scheduled sampling, study participants who identified in their questionnaire a life event which could have changed their virome for a given sample, or other unforeseen variance in sample collection.

An assembly independent method was also employed using the identified stable viral families for further refinement of viral taxonomic identification at the species and genus level and investigation into putative human identification markers. Based on the preliminary taxonomic assessment of the most abundant and stable viruses, the corresponding reference sequences belonging to the viral families Papillomaviridae (n = 6546 genomes), Genomoviridae (n = 881 genomes), Baculoviridae (n = 1325 genomes), as well as the order Caudovirales (n = 37,087 genomes) were downloaded from the NCBI nucleotide database for subsequent analysis. The resulting reference database generated contained a total of 45,829 reference sequences. Trimmed and quality filtered sequencing reads were mapped to these reference sequences using Bowtie 2 v.2.3.5 [31]. The resulting mapped read counts to database reference sequences were acquired by filtering with SAMtools

[32]. The read counts for each mapped reference genome (NCBI reference genomes) and putative viral genome assemblies (identified in this study) were utilized for further investigation of viral diversity and persistence of selected viral families and for statistical evaluation using R. To assess stability of the identified markers a Jaccard dissimilarity matrix was produced with the option of binary set to true, with the resulting distance of 0 being all classes being the same and 1 being all classes being disjointed. The subsequent Jaccard distances were averaged between a subject's sample and all other samples (intersubject dissimilarity) and averaged all distances between a subject's sample and all other samples collected from that same subject (intrasubject dissimilarity). We have focused on Jaccard in this portion of the study because it is more discriminating for samples consisting of presence-absence data [36]. A series of pairwise Wilcoxon rank sum tests were performed between the intrasubject dissimilarity distances and the intersubject dissimilarity distances to assess stability of the markers.

2.9. Viral metagenome diversity

Alpha diversity (α-diversity) was assessed using the Shannon alpha diversity metric [37]. To evaluate if the contig diversity of a subject significantly changed across body sites and time, a one-way ANOVA using repeated measures was used. For Beta diversity (β-diversity), a binary Jaccard dissimilarity matrix with the option of binary set to true was generated based on presence and absence of identified putative viral human identification markers. Beta diversity was visualized using principal coordinate analysis (PCoA). The subsequent binary Jaccard dissimilarity matrix was used for PERMANOVA analysis using the Adonis test in the Vegan package v.2.5–7 [38] to assess changes of the virome between subjects. These analyses were all conducted using the R statistical package and using the R associated tools Phyloseq and Vegan [33,34,38]. All metadata, list of markers, contig sequences, annotation files, and scripts described in the material and methods are publicly available and archived at: https://github.com/HerrLab/Graham_2021_forensics_human_virome.

3. Results

3.1. Virome assembly and annotation

Skin virome samples were collected from 42 individuals across three locations (left hand, right hand, and scalp) over a six-month period for each subject - with samples representing the initial timepoint (day 0), 2-weeks, 1-month, 3-month, and 6-month time points (scheduled from the initial sampling date). Anatomical sampling locations of hands and scalps were chosen based on their potential to be of forensic relevance, as well as their potential to provide a high level of viral diversity and allow stability as shown in our pilot sampling and previously published work [15,39].

Samples were pre-processed using 0.2-micron filtration to remove eukaryotic and prokaryotic cells prior to high-throughput sequencing. After sequencing, we processed the sequencing reads through bioinformatic analyses to remove human genomic contamination by removing any reads that mapped to the hg19 human reference genome. Additionally, to ensure that the viral contigs contained negligible levels of bacterial contamination, sample sequences were evaluated for the presence of 16S rRNA genes. Samples contained on average 0.002% of ribosomal reads per sample and a maximum of 0.16% of ribosomal reads. Previous studies have demonstrated that if viral metagenomes have less than 0.2% 16S rRNA reads, these datasets are enriched for viral sequences and have minimal and likely negligible bacterial contamination [40]. As such, the low occurrence of bacterial 16S rRNA sequences indicated that the resulting virome dataset was adequately enriched for human-associated viruses with minimal contamination. Subsequent quality filtered reads were used to evaluate viral diversity and viral stability with the proof-of-concept goal of identifying viral markers for human identification.

Out of the read assemblies, contigs larger than 1000 bp were only considered for subsequent analysis. The value of 1000 bp in length is a standard value employed in numerous bioinformatics packages used in this study [23,26,28]. In total, out of the 952,760 contigs, 62,101 contigs were $>$ 1000 bp and thus retained for further analysis. This resulted in the contigs having a N50 value of 1970 and the longest contig length being 54,929 bp. To further identify contigs of viral origin based on currently available viral sequence information, contigs were analyzed using CheckV v.0.7.0 [26]. Out of the 62,101 assembled contigs, 1400 were identified as having a known viral gene and assumed to be truly viral in origin based on similarity to current databases [26]. Of the 1400 contigs with a gene corresponding to viral databases, we removed 102 contigs for having sequence similarity to either human or animal (e.g., *Canis lupus*, *Felis domesticus*, etc.) genomes or were considered to be a prophage, which resulted in 1298 final contigs. Due to the lack of large-scale studies evaluating the diversity of the human skin virome as well as poor reference databases focused on viral diversity, we estimate that a portion of the filtered and removed contigs may contain novel viruses that were not able to be annotated using current methods and reference databases.

3.2. Human skin viral diversity

The 1298 identified contigs were taxonomically classified using current viral databases and viral annotation bioinformatic tools. The distribution of abundance by taxonomy of the contigs across all samples is represented in Fig. 1. A majority of the identified viral contigs could not be fully annotated using currently available viral reference databases and existing bioinformatic tools. Even the contigs that were identified as viral through CheckV, could not be classified to lower taxonomic levels using currently known reference viral genomes suggesting that existing public viral databases are poorly representative of human skin viral diversity. This paucity of viral reference material contributed to taxonomic uncertainty for many contigs in this study. Due to the fact that we sequenced the DNA virome, and did not attempt to sequence RNA viruses, the contigs expectedly show homology to double and single stranded DNA viruses. However, many of the DNA viruses identified were not able to be taxonomically classified due to a lack of sequence representation in the current viral databases. Many of these viruses were identified as highly abundant in the core human skin virome (Fig. 1). Among the double stranded DNA viruses identified, the viral order Caudovirales was the most abundant order detected (Supplementary Fig. 2). This is not surprising, as there is a disproportionate amount of Caudovirales viral genomes available in viral reference databases. As for the identified single stranded DNA viruses, highlighted in Supplementary Fig. 3, the most abundant taxa were that of small circular DNA viruses which included Papillomaviruses and Cress-like DNA viruses. Papillomaviruses and Polyomaviruses are common skin associated opportunistic pathogens and the identification of Papillomavirus genomes was expected [15,41].

3.3. Subject viral diversity abundance across skin sites

The top ten most abundant viral families identified for each location across the entire group of participants are shown in Fig. 2. Of the most abundant viral families observed, Papillomaviridae and varying viral families that belong to the order Caudovirales were most abundant. This is consistent with the overall virome diversity identified across all individuals (Fig. 1). As seen in Fig. 2, the virome was different from subject to subject showing the discriminatory capability of the virome that could be used for human identification. When comparing across skin site locations within an individual, in some instances there was either a complete absence or addition of highly abundant viral families. One such family was the *Streptococcus* satellite phage Javan 305, although

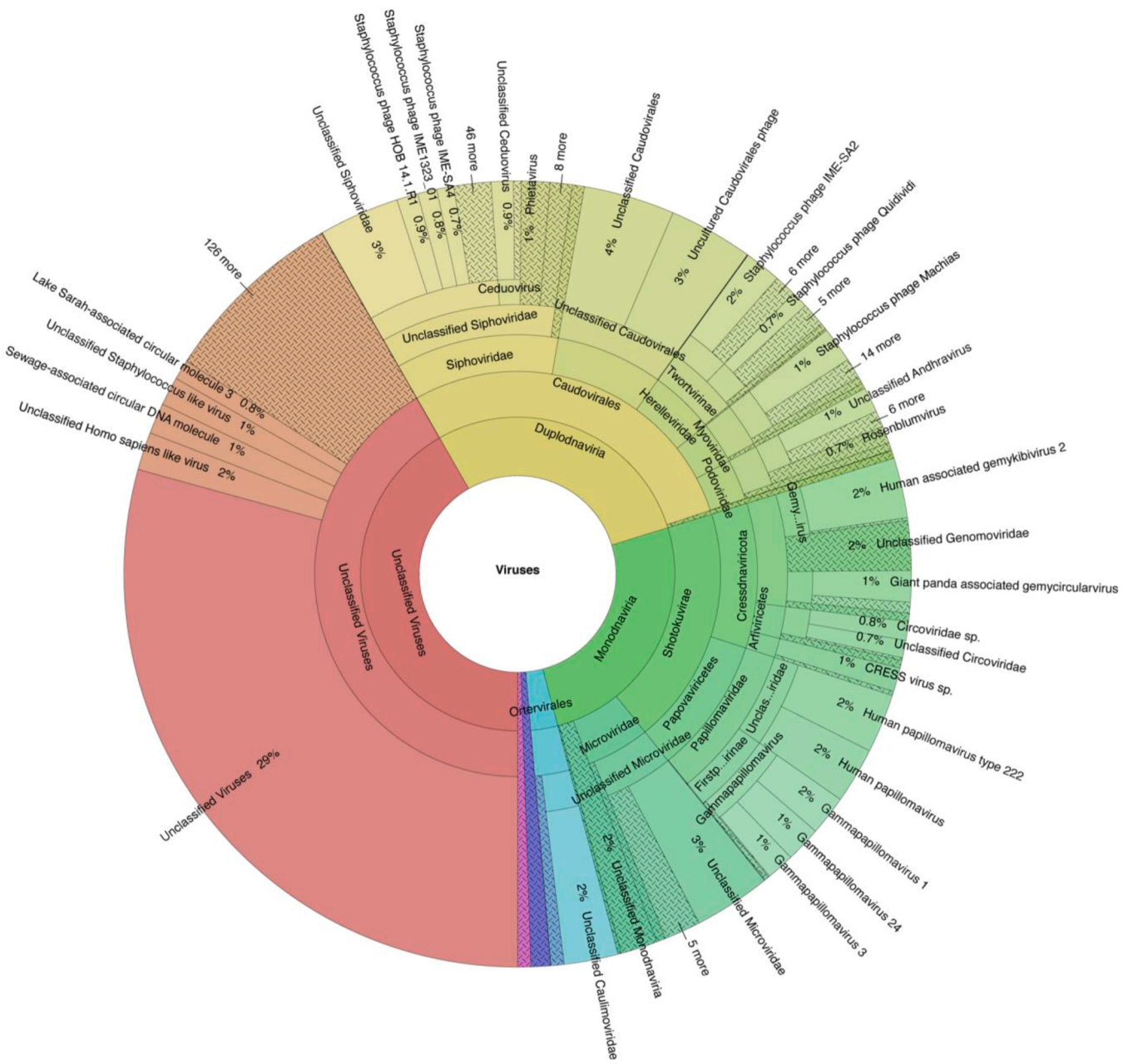

Fig. 1. Overall distribution of viral taxonomy identified from 652 human skin virome metagenome samples derived from 42 subjects. Distribution consists of the total abundance of each taxonomic classification assigned to the assembled contigs > 1000 bp from human skin metagenome samples that were identified as containing at least one viral gene using CheckV v.0.7.0 [26]. Double and single stranded DNA viruses composed more than 50% of the overall diversity abundance of the identified skin virome, however a large portion of highly abundant taxa were unable to be classified.

most of these other viruses have yet to be isolated or characterized. When comparing the relative abundance of viral families, we observed clear differences across individuals. Considering the most abundant viral families, one could draw clear distinctions between individuals, while still maintaining higher levels of similarity within an individual across multiple skin collection locations. This observation was statistically assessed by comparing the family level community differences between subjects versus within subjects across locations. It was found that between subjects was significantly more different in viral family level diversity than within subjects across skin site sampling locations ($P < 2.22 \times 10–16$) (Supplementary Fig. 4). These findings further support the notion that the human skin virome is an individual characteristic and potentially useful for discrimination between different subjects.

3.4. Stability of the skin virome over time

We recognize that some specific viruses may be singularly diagnostic (for example, we identified contact with cats by traces of feline leukemia virus or feline papillomavirus on their hands), but our goal in this proof-of-concept study was to characterize viral markers for human identification and as forensic evidentiary material. In order to do this, markers needed to be identified with accuracy and consistency over time allowing for comparison of viral profiles collected at different time points. Therefore, the persistence of contigs we identified were evaluated across individuals and locations within an individual over time. As previously mentioned, stable viral taxa were considered by the presence of a viral contig or taxa in at least four out of the five time points for a given individual at a certain skin location. This analysis identified the

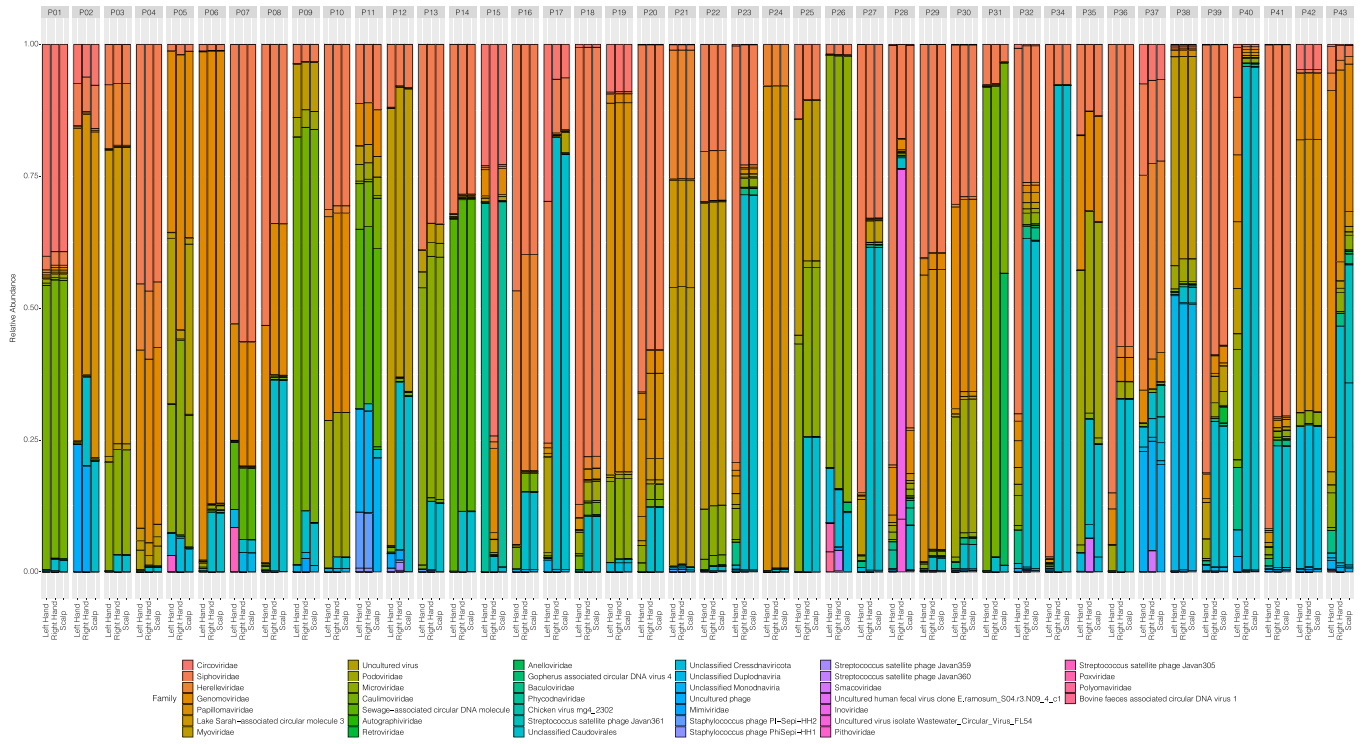

Fig. 2. Relative abundance of the top ten most abundant identified viral families per subject by location shows intrasubject similarity, with increased intersubject differences in diversity. Of the identified viral contigs, abundance of all taxonomically assigned viral families were summed across all time points by anatomical sampling location within a subject. Each bar represents a location within a subject, as indicated at the bottom. The bars are then further separated by subject as indicated by subject notation at the top. All contigs that were not able to be taxonomically classified at the family level using current databases were not included. Clear similarities in taxonomic abundance across locations within an individual can be observed. However, when comparing across subjects (intersubject), there is increased dissimilarity in relative abundance taxonomic profiles than that of within subject (intrasubject) comparison across locations.

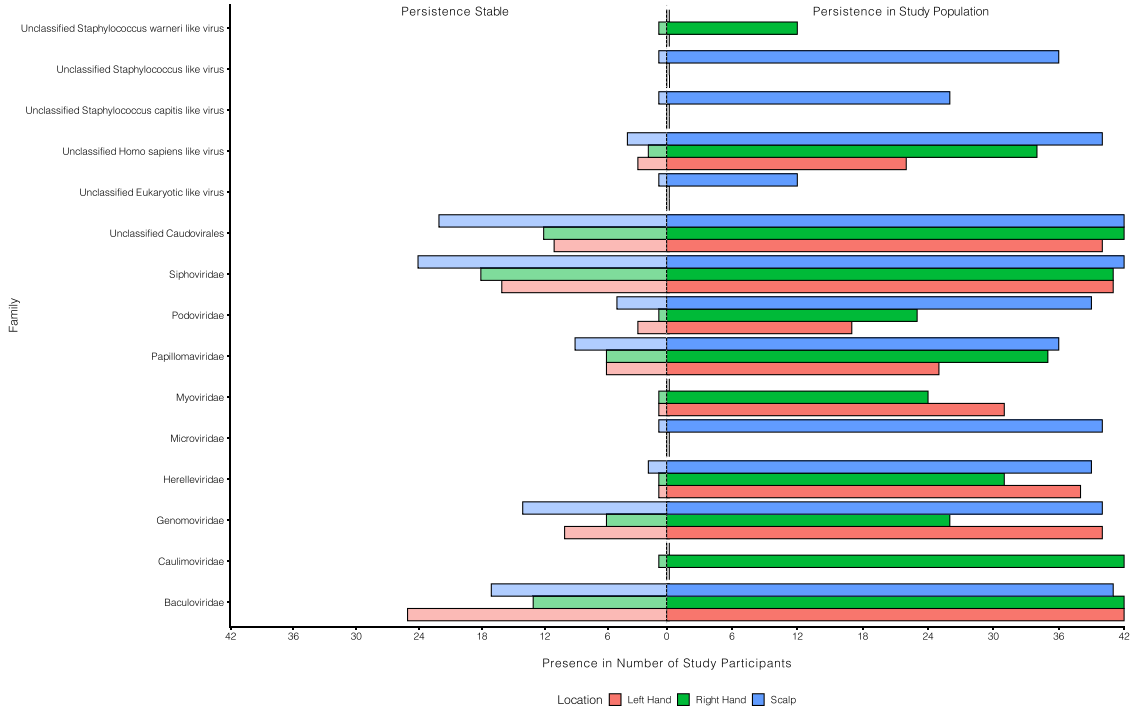

Fig. 3. Persistence of stable identified viral families within the study population. Bars represent the number of study participants (42 total) within the study population where a denoted viral family was stable (left side; persistence stable) and the overall presence regardless of stability (right side; persistence in study population). Prevalence for each identified viral family was determined across the five sampled time points of initial sampling (day 0) and 2-week, 1-month, 3-months, and 6-months post initial sampling. Viral families that were present in four out of the five time points at an anatomical location within a subject were considered stable. Only viral families that exhibited stability in at least one individual within the population are displayed.

viral families, Baculoviridae, Genomoviridae, Herelleviridae, Myoviridae, Papillomaviridae, Podoviridae, Siphoviridae, Unclassified Caudovirales, and Unclassified *Homo sapiens* like virus to be stable in at least one individual across all three skin locations and thus were considered as potential target viral families for further investigation to develop biological markers for human identification (Fig. 3). Although these viral families may have stability in certain individuals, they may be absent or temporally transient in other individuals, therefore exhibiting variation in virome stability across subjects.

3.5. Identification of putative viral skin markers for human identification

Due to high occurrences of temporal transiency of the identified viral families within certain subjects, categorizing the contigs at the family level alone is not suitable as a taxonomic marker for human identification and a higher level of annotation should be used (e.g., such as that of genus or species or unique contigs) (Fig. 3). Of the classified contigs, marker identification and stability were therefore not only assessed at a family level but also at the level of viral genus and species resulting in identification of eight left hand, 11 right hand, and 23 scalp associated stable viral species (Table 1, Set A; Fig. 4A).

To improve viral genome recovery and reduce metagenome assembly bias, trimmed and quality filtered reads were mapped to all NCBI nucleotide viral reference sequences associated with target viral families identified to be stable across all three skin locations. All NCBI nucleotide viral sequences associated with the order of Caudovirales were used for this analysis due to the fact that several of the target families, such as Siphoviridae, Podoviridae, and Myoviridae, all fall under this order. Mapped counts to the reference genomes were evaluated similarly to that of classified viral contigs with species level analysis as previously described. In addition, to identify candidate species or groups of species for human identification, virome stability within an individual over time was used as an experimental factor. Again, viral taxa found to be present in four out of the five time points within a location for at least one individual were considered to be stable. In total, 46 left hand, 56 right hand, and 81 scalp associated stable viral species, using assembly independent and database dependent mapping, were identified as putative viral markers (Table 1, Set B; Fig. 4B).

To address uncharacterized viruses that were unable to be annotated as well as viral identification bias based on our reference databases which only used known viral genes, we assessed the stability within a subject using all contigs that were not identified as containing at least one viral gene using CheckV as previously described. The subsequent contigs identified within a subject as being stable over the course of sampling were retained. As before, using NCBI's Blastn algorithm we annotated these contig sequences. Contigs containing $< 70\%$ identity to a known organism or contigs with open reading frames with no sequence similarity to prokaryotic or eukaryotic genes were considered to be putative viral sequences, although viral origin could not be fully confirmed. For this independent putative viral marker identification, 29 left hand, 34 right hand, and 65 scalp contigs were identified as being stable (Table 1, Set C; Fig. 4C).

To address database bias, uncharacterized viral taxa, and metagenomic assembly error, the three sets of putative viral markers for human identification were compiled. When all three sets were taken into account a total of 188 non-redundant viral species or contig markers were identified as stable and are thus proposed as potential viral markers for human identification (Fig. 4D). To further limit the number of markers, only markers that were found to be stable across all three anatomical locations were retained for marker identification, resulting in a final set of 59 putative human identification viral markers (Fig. 4D). A heatmap, shown in Fig. 5, was rendered to visualize marker persistence profiles of the 59 viral markers across the five time points by subject by location. Of the viral markers, seven markers (*Staphylococcus* phage vB_SauH_DELF3, Unclassified Baculoviridae, *Escherichia* virus Lambda, *Autographa californica* multiple nucleopolyhendrovirus 1, *Streptococcus* phage phi-SC181, Marine virus AFVG_25M557, and *Streptococcus* phage phiJH1301–2) were stable and present across all individuals, whereas all other markers retained diagnostic and/or discriminatory power across individuals.

Table 1
Description and quantity of the three human skin viral biomarker sets for human identification.

	Marker Identification Method		Number of Stable Viral Species/Contigs			
	Assembly[a]	Database[b]	Left Hand	Right Hand	Scalp	Description
Set A	Dependent	Dependent	8	11	23	Species level Check V identified and database classified viral contigs
Set B	Independent	Dependent	46	56	81	Species level NCBI reference database sequence reads
Set C	Dependent	Independent	29	34	65	Viral contigs with no taxonomic classification
Total Non-Overlapping Markers[c]			80	97	161	

[a] Assembly bioinformatic tool used in this study was MEGAHIT v.1.2.8 [23].
[b] Databases used for annotation were current databases associated with nucleotide-based classification tools, such as Kraken2 v.2.0.8-beta [27], Demovir [28], and Blastn [29], and the classification tool based on protein coding sequences, Kaiju v.1.7 [30].
[c] Comprised of all non-overlapping (i.e., unique) markers from all three sets.

3.6. Statistical assessment of identified viral skin markers for human identification

Three sets of putative markers for human identification were identified as previously described (Table 1, Set A, Set B, and Set C). These markers were chosen on their basis for stability across all sampled skin site locations within at least one individual thus substantiating their stability within one individual in the study population. However, in order to test the viability of the identified markers, the stability of the markers across the entire population and their differentiation across individuals was evaluated. To do so, for each marker set (which we labeled as "Set A", "Set B", "Set C", and "Overall" which included all three sets combined) a binary Jaccard dissimilarity matrix was produced to compare intrasubject versus intersubject variation on the basis of presence or absence of each of the markers within each set. For all three sets, intrasubject variation was significantly less dissimilar than that of the intersubject variation (Set A: $P = 0.00011$; Set B: $P = 4.4 \times 10^{-10}$; Set C: $P = 1.5 \times 10^{-10}$) (Fig. 6A–C). In order to evaluate the sets in combination with one another, we compared intrasubject variation and intersubject variation using all markers for each skin site location. We additionally compared an overall set which encompassed all locations and subjects. For all site locations and the overall comparison (all locations and overall marker set comparison), there was a highly significant difference between the intrasubject variation and the intersubject variation (Left hand: $P = 6.3 \times 10^{-14}$; Right hand: $P < 2.22 \times 10^{-16}$; Scalp: $P = 3.6 \times 10^{-10}$; Overall: $P = 5.3 \times 10^{-15}$) (Fig. 6D–G). Thus, showing that the marker sets, on the basis of presence or absence, are significantly more similar within an individual across time points than compared to the base-line presence or absence of those same markers in the rest of the subject population.

In order to evaluate subject differentiation across the population, the differences in marker diversity, using presence and absence of the

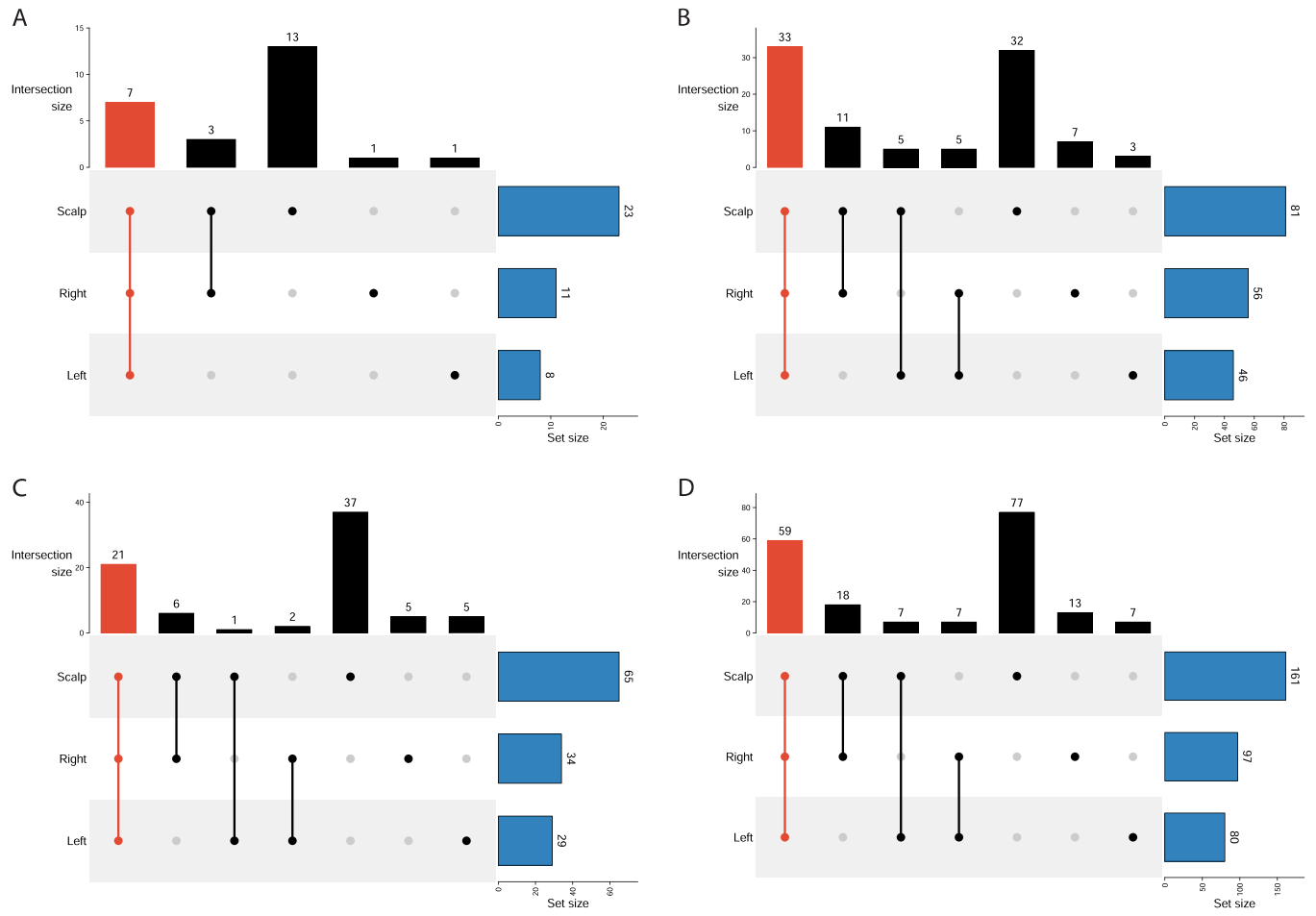

Fig. 4. Prevalence of marker stability within the marker sets across the three anatomical locations. Plots show the distribution of viral markers that are stable in the study population across the three anatomical locations for Set A (A), Set B (B), Set C (C), and all non-overlapping markers from the three sets combined together (D). Stability was defined as the marker being present in four out of the five time points within at least one subject at the denoted location. Highlighted in red is the quantity of markers that were deemed to be stable in all three anatomical locations and thus used for downstream evaluation of population marker profile comparisons. Bars on top (Intersection Size) represent the number of markers that were considered to be stable at the combination of locations as noted by the scatter plot below the bar. Bars shown in blue (Set size) are the overall number of markers within the set that were found to be stable at each anatomical location. The 59 markers that were deemed to be stable across all three locations from all three sets combined (D), are considered to be the putative human virome profile markers for human identification. Set composition is described in Table 1.

identified markers, was evaluated. We compared within sample diversity (α-diversity) of all of the markers using the Shannon index as a diversity metric (Fig. 7A). Using an ANOVA test, we found that there was a significant difference in the amount of marker α-diversity across subjects ($P = 0.002$) and a slight significance difference across sex ($P = 0.02$). However, there was not a significant difference in α-diversity across the three locations within subject ($P = 0.066$). To establish if between subject-to-subject diversity (β-diversity) was significant, the Adonis test in the R package Vegan v.2.5–7 [38] was used to run PERMANOVA tests using the Jaccard distance method in order to evaluate differences on the basis of presence and absence and not abundance (Fig. 7B). It was found that for β-diversity there was a significant difference across subjects ($P < 0.001$; $R^2 = 0.3415$).

4. Discussion

The utilization of alternative sources of biological data for forensic human identification should incorporate aspects of both target marker stability over time and the ability to utilize those markers to have probative discriminatory power across individuals within a population. This study set out to address the viability of the human skin viral microbiome as a proof-of-concept for its stability and utilization as a source for human identification. Previous studies, such as those by [6,7,10], have identified a suite of bacterial biomarkers (along with one bacteriophage) for human identification. Additionally, studies have shown certain viral taxa associated with human skin disease have yearlong stability and thus could be an additional microbial target for human identification [42]. However, no previously published studies have exclusively investigated the potential utilization of the skin virome for forensics or have identified a panel of potential taxonomic and sequence variant viral markers for human identification. The goal of the study presented here was to test the feasibility and proof-of-concept of targeting the viral component of the human skin microbiome, as has been done for the bacterial microbiome, for forensic purposes.

In this study, we sequenced a total of 652 human skin viral metagenomes and analyzed that data to identify viral markers for human identification. Samples were pre- and post-sequence processed to reduce eukaryotic and prokaryotic contamination. Bacterial contamination was assessed by quantifying the presence of 16S rRNA genes in the viral assemblies. Samples contained a low amount of bacterial 16S rRNA sequences which suggested our viromes to be highly enriched for human-associated viruses with minimal contamination. Enrichment for virus-like particles (VLPs) as opposed to an overall bulk shotgun microbiome approach allows for deeper sequencing of viral genomes which aids in increased viral sequence data recovery and viral annotation [43]. A study conducted by Trubl et al. [43] found that enrichment for VLP

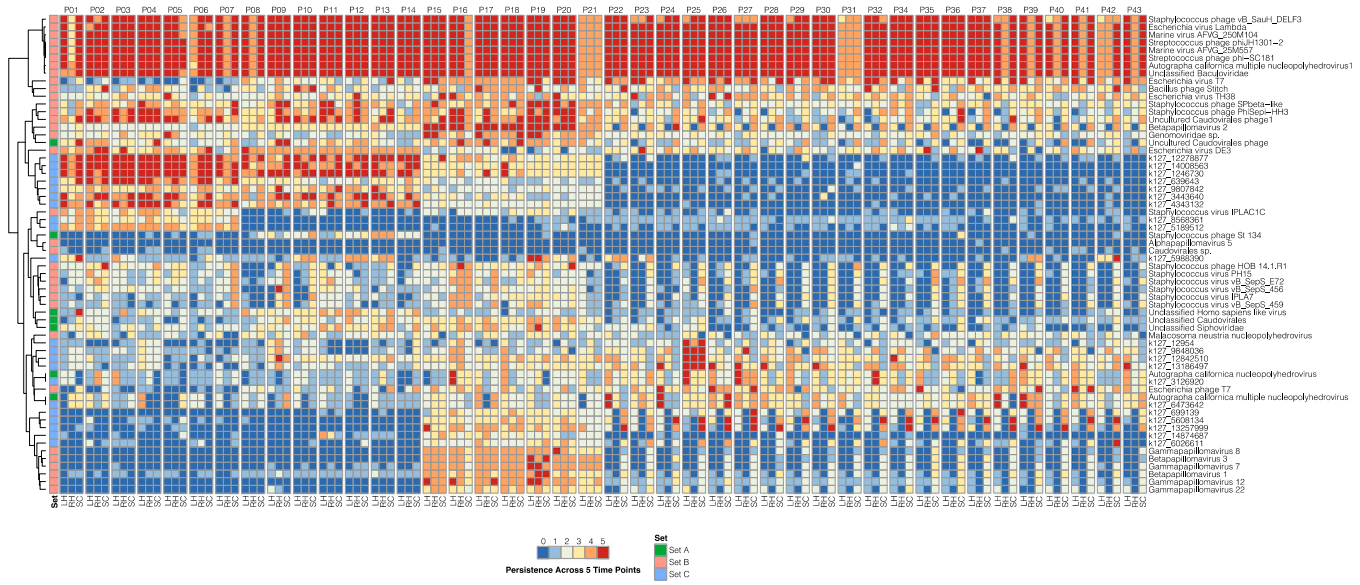

Fig. 5. Persistence of the 59 markers over time across the subjects and locations shows distinct profiles for each individual. Profile heatmap where each column is an anatomical location (LH = Left Hand; RH = Right Hand; SC = Scalp) which is separated by subject as denoted on top. Each row is associated with an identified viral stable marker from the overall marker set with all three sets (e.g., Set A, Set B, and Set C) combined. Rows are clustered based on marker prevalence similarity across the samples. Persistence of the marker in the five time points is represented by the color scale with red being present in five out of the five time points and blue being present in zero out of the five time points. The top seven markers in the heatmap were found to be stable (stability defined as prevalent in four out the five time points within a location within an individual) across all subjects and thus proposed for future studies looking at genetic variation within that viral species for subject discrimination within the population.

metagenomes outperformed bulk metagenomes 2-fold in overall viral recovery. An additional study conducted by Gregory et al. [44] found that the use of a bulk metagenomic methodology resulted in biased recovery of viruses and prophage that were actively infecting bacteria where enrichment for VLPs resulted in greater amounts of free viral particles. For the proof-of-concept of this study, we provided alternative biomarkers from previous studies (i.e. human or bacterial communities) just using a bulk metagenomic methodology with laboratory filtering and bioinformatic data filtering methods. In the context of forensics, intuitively one would want to target free viral particles seeing as how these particles have the greater potential to be transferred from a subject to an evidentiary object whereas targeting prophages implies the necessity of host-specific bacteria within the collected microbiome sample. By utilizing viral enrichment methods, we were able to identify and recover eukaryotic and human infecting viruses which will not be affected by as many environmental factors such as that of antibiotic drugs or hand washing using antibacterial soap. Additionally, by targeting human-associated viruses this will allow for the generation of additional biomarkers for instances when bacterial biomarkers are affected or cannot be used such as that of the environmental effects previously mentioned.

The sequence data isolated from 42 study participants across five time points for 3 sampling locations on the human body was assembled. Of the 62,101 assembled contigs with >1000 bp, 1298 contigs were identified as viral based on current database dependent identification tools. This is a small percentage of the overall metagenome assembly. Due to the fact that annotated putatively viral genes within the assembled contigs were taxonomically identified by comparing to current viral databases, our data indicated that these reference databases are lacking in viral diversity and this lack of representation makes it difficult to study viromes in any environment. Thus, similar studies coupled with genome annotations are greatly needed to improve viral annotation. As such, this study should be viewed as a first step towards uncovering viral diversity and stability in the human virome in addition to its application to forensics.

Of the identified viral contigs, both double and single stranded DNA viruses were identified. The most abundant double stranded DNA viruses detected were those of bacteriophage belonging to the viral taxonomic order Caudovirales. Previous studies have found the skin virome diversity to be mainly composed of Caudovirales bacteriophage [15,42] and our findings here were in agreement with this observation. The viral families identified in this study that fall under the order Caudovirales included Siphoviridae, Podoviridae, and Herelleviridae. These viral families include bacteriophage that are obligately associated with bacteria commonly found in the skin microbiome (i.e., *Staphylococcus* and *Streptococcus*) [15,45,46]. With bacteria being the dominant component in the skin microbiome, it is logical that the abundance of bacteriophages would show a similar level of abundance in the human skin virome. These bacteriophages may help control bacterial populations on the skin and may help structure the bacterial communities of the human skin microbiome with relation to environmental inputs (seasonality, humidity, etc.) and lifestyle of the person (travel, activities, etc.). In addition to double stranded DNA viruses, a large amount of single stranded DNA viruses were identified in the skin virome samples. Small circular DNA viruses, those that are typically associated with eukaryotes, such as those found in the viral families of the Adenoviridae, Anelloviridae, Circoviridae, Herpesviridae, Papillomaviridae, and Polyomaviridae, have all previously been reported to be associated with the human skin virome. A large proportion of the single stranded viruses identified are Papillomaviruses, which are common skin associated viruses that may act as opportunistic pathogens [47]. We also observed a high abundance of the small circular single-stranded DNA viruses belonging to the phylum of Cressdnaviricota across all subjects in this study (Fig. 1, Supplementary Fig. 3). These viruses have previously not been reported at high abundance in the human skin virome. The lack of Cressdnaviricota virus reporting in previous human skin virome studies may be attributed to the recent discovery of novel Cress-DNA viruses and their recent addition to NCBI databases [15,41,45,48]. Of the contigs that were annotated under *Cressdnaviricota* many of them only displayed minimal similarity to reference Cressdnaviricota viruses suggesting we have identified a group that is more diverse than previously reported. These viruses did contain a *Rep* gene which is specific to

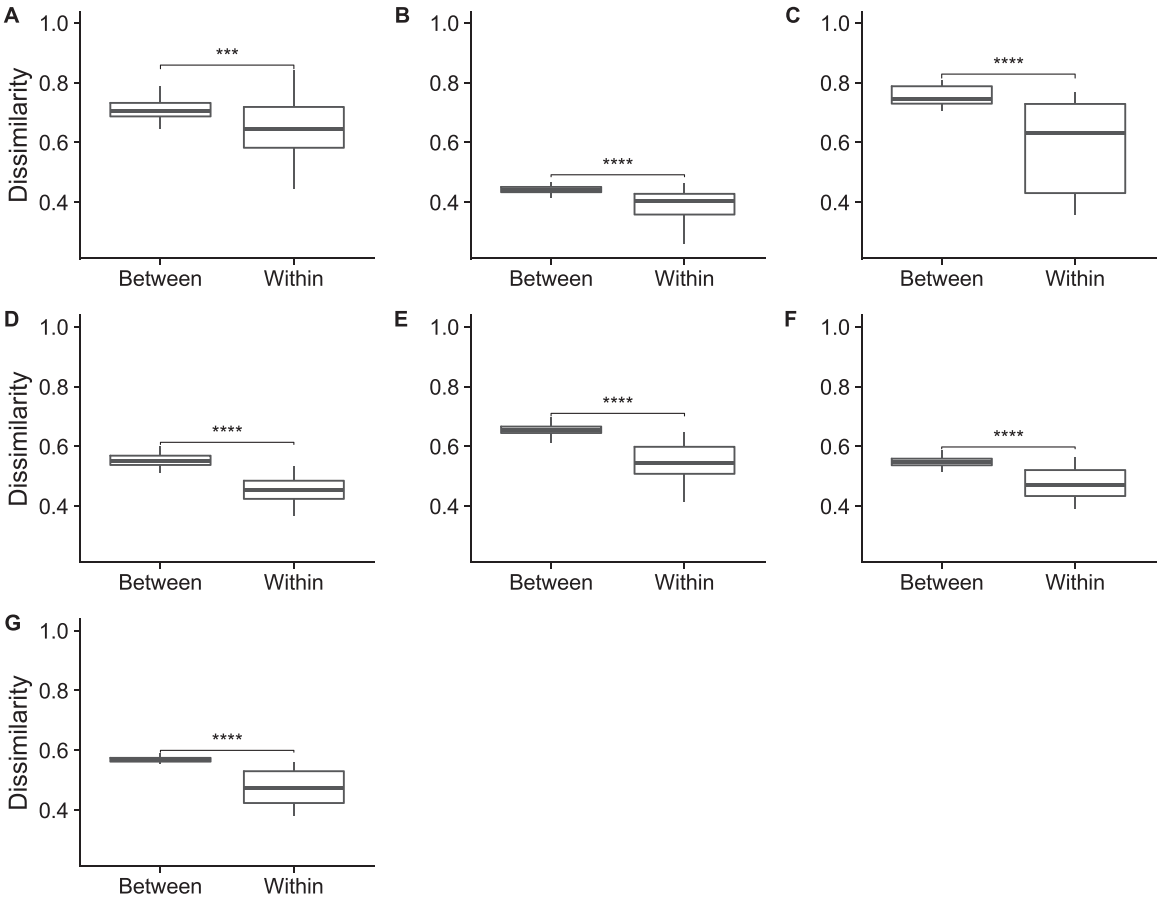

Fig. 6. Profiles of identified viral biomarkers are stable over time as indicated by a significant decrease in intrasubject dissimilarity as compared to intersubject dissimilarity. Boxplots of within-subject (comparison of samples within subject across time) and between (comparison of sample to all other subject samples) subject binary Jaccard dissimilarity distances; 0 = samples are identical, 1 = samples are disjoint. Statistical comparison of intersubject (between-subject) dissimilarity and intrasubject (within-subject) dissimilarity was done by Wilcoxon ranked sum tests. Intersubject marker presence and absence profile was significantly more different than within-subject across time for the identified marker Set A (***; $P < 0.00011$; $n = 42$ subjects) (A), Set B (****; $P < 4.4 \times 10^{-10}$; $n = 42$ subjects) (B), and Set C (****; $P < 1.5 \times 10^{-10}$; $n = 42$ subjects) (C). When all non-overlapping identified stable markers from the three sets were combined, the between-subject marker presence and absence profile was significantly more different than within-subject across time for the anatomical locations left hand (****; $P < 3.6 \times 10^{-10}$; $n = 42$ subjects) (D), right hand (****; $P < 2.2 \times 10^{-16}$; $n = 42$ subjects) (E), scalp (****; $P < 6.3 \times 10^{-14}$; $n = 42$ subjects) (F), and overall all locations combined (****; $P < 5.3 \times 10^{-15}$; $n = 42$ subjects)(G). Description of composition of marker Set A, Set B, and Set C are described in Table 1. * significant at $P < 0.05$; ** significant at $P < 0.01$; *** significant at $P < 0.001$; **** significant at $P < 0.0001$.

the phylum Cressdnaviricota, however they had low sequence similarity to the sequences in the current database suggesting the viruses identified in this study could be novel viruses related to Cressdnaviricota and other small DNA viruses.

Many contigs that were identified as being of viral origin were small and circular in nature. These viruses have similarity to other small circular DNA viruses in current viral reference databases. However, due to their low percent identity to any known virus or organism they remained unclassified in our dataset at lower taxonomic levels (e.g., at the level of viral genus or species). This was especially evident in the unclassified viruses having similarity to viruses belonging to family Microviridae. As previously mentioned, the skin virome contains many novel viruses, particularly noted here in the family Microviridae. In order to better characterize the viral diversity in the human skin virome, more research into viral discovery is needed. Since there was a large proportion of viral contigs having no similarity to characterized viruses, we implemented a database independent approach paired with database dependent methods to alleviate annotation inaccuracies and missing reference material.

To evaluate the applicability of the human virome for subject identification, viral contig diversity and abundance across locations and subjects was further analyzed to identify the top ten most abundant viral families for each subject by sampling location (Fig. 2). Distinct differences in the relative abundance of these contigs were observed across individuals. Relative abundance across locations within an individual was also visualized in Fig. 2 and similarity across all three locations was observed within most individuals. However, differences were noted for some subjects between locations (right hand, left hand, and scalp). It is possible that this may have to do with hand dominance coming into contact with that individual's hair and scalp thus sharing similar viral taxa across locations. The main difference observed across these locations is either the addition of a viral family or the loss of one.

The relative abundance of the most abundant annotated contigs visually displayed greater variation between subjects compared to that of within subject variation across sampling locations (Fig. 2). This suggests viral marker presence can be discriminatory for differentiation of individuals within the study population. It is important to note, however, that Fig. 2 is a comparison of only subject and location within subject, with time of sampling and viral stability not taken into account. In Fig. 2, we evaluated the top ten most abundant taxa for each sampling anatomical location and subject summed across the five collection time points. Therefore, this comparison should not be used as an ultimate viral marker set for human identification purposes and thus further evaluation into viral stability was conducted.

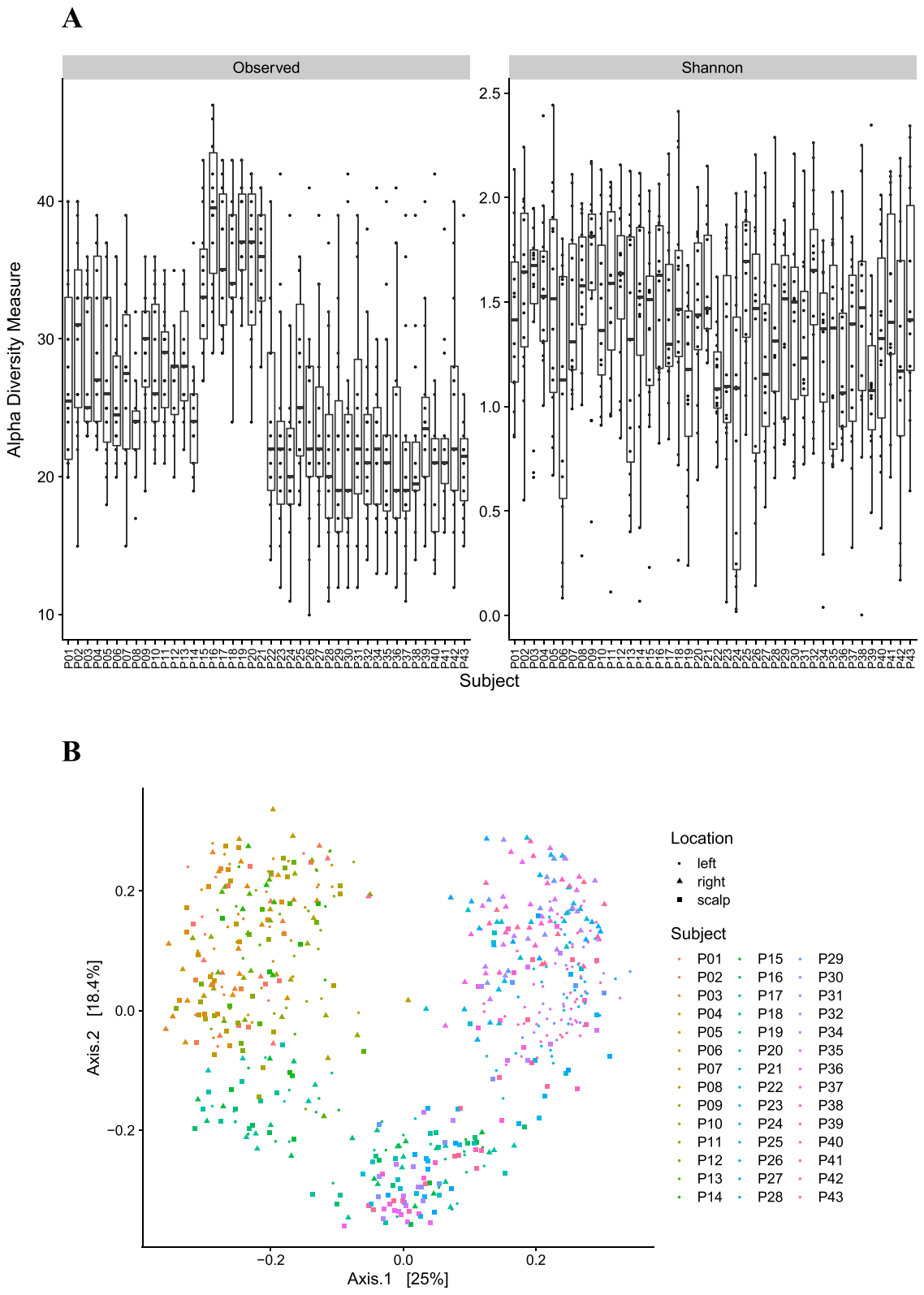

Fig. 7. Identified viral marker diversity across subjects. (A) Boxplots comparing α-diversity of the identified viral markers in subjects using the α-diversity measure Shannon [37], as well as the uncorrected number of observed markers within subject. An ANOVA for the variables of subject ($P = 0.002$) and gender ($P = 0.02$) were found to be significant, while location within subject was not ($P = 0.066$). (B) PCoA plot of binary Jaccard dissimilarity distances of virome samples by subject to assess β-diversity of the identified viral markers across subjects. Clustering by subject was significant ($P < 0.001$; $R^2 = 0.3415$) by PERMANOVA analysis using the Adonis test.

Not only do viral markers need to be diagnostic for individuals, but they also need to be stable over time. Therefore, for forensic purposes, we recognize that the core stable virome needs to be targeted for biomarker development. The virome, especially those viruses found on body locations that are consistently exposed and interact with objects and the environment, will potentially have a large temporal component in addition to the stable component [15,39]. Temporal viruses do have the potential to provide informative forensic information that would be unique to certain lifestyle characteristics like occupation, contact with animals, recent travel, or hobbies (such as gardening, etc.). Temporal viruses, while not stable for direct forensic applications, could provide important information towards identifying circumstantial characteristics specific to an individual. However, for the purpose of acting as an alternative genetic source to traditional STR methods, target viral markers should be stable over time, and therefore, we required that viral markers be consistently present within an individual's skin virome during our sample collecting regime, which we repeated for each individual at all three locations over a time course of six months.

For the purposes of this study, we considered a virus to be stable if a particular consistently annotated virus was present in at least four out of the five time points collected within a single anatomical location for an individual. Of the viral families identified in the overall assembly, 15 families were considered to have stability in at least one location within an individual. Of the 15 families, nine viral families presented stability in at least one individual for all three body location sites. Of note, many of the nine families fell under the order of Caudovirales. As previously mentioned, several studies have evaluated and classified the core bacterial microbiome on the skin and thus it is not surprising that bacteriophage would be both present and temporally stable seeing as how their bacterial hosts are also present and stable on human skin [39,45,49]. In a few older studies, Papillomaviruses were found to not be stable or not less stable as the communities of certain phage viruses as well as both bacterial and fungal microbiome communities [15,39]. However, in this study, in addition to families of Caudovirales, the family of Papillomaviridae was found to be stable which is consistent with findings observed in another study [42]. Interestingly both Baculoviridae and Genomoviridae were also found to be stable across multiple individuals in all three physical locations sampled in this study. Notably, *Baculoviridae* are sometimes used for bioengineering purposes in sequencing laboratories so there is a chance that their presence in our study is due to contamination from a sequencing provider, however, we think this is highly unlikely as certain Baculoviridae have the ability to infect mammalian cells, such as *Autographa californica* multi-nuclear polyhedrosis virus. Therefore, our identification of similar viruses may potentially be a true biological observation that has not previously been accounted for due to other studies assuming it was contamination [50]. Although Genomoviridae have previously been identified from human vaginal samples, very few studies have explored these small single stranded DNA viruses and their diversity and stable presence as a part of the human skin core virome [51]. Interestingly, previous studies have identified Genomoviridae as being a stable and persistent virus in other mammals such as that of bats [52].

Of the nine viral families that we identified as being temporally stable, their continuously observed persistence was found across many, though not all, of the subjects in our study (Fig. 3). This is potentially due to certain viral genera or species within a particular family being temporally stable as opposed to similar genera or species within that same family that may not have been temporally stable. Additionally, there is a possibility that our viral metagenome assemblies could be biased due to the high amount of repetitive sequences seen in some particular viruses. Therefore, in order to reduce assembly-based bias and increase annotation, we trimmed sample reads and mapped that data to all genome sequences available in NCBI's nucleotide database pertaining to the order Caudovirales and the viral families Papillomaviridae, Genomoviridae, and Baculoviridae, to improve viral detection and sample annotation. Species level mapped NCBI reference genomes were evaluated during this process. In addition to stability of reference mapped counts, species level stability was also assessed for viral contigs. Species that were identified as being persistent across all three locations in at least four out of the five time points for an individual were considered for putative viral markers for human identification.

In order to address viral diversity missing from databases, we first filtered sequencing reads on the basis of base-call quality and then mapped the sequencing reads to the total contigs in our viral metagenome assembly. Contig sequences that were stable within an individual for each sampling location were analyzed with Blast using the Blastn alignment algorithm. Contigs that did not share a high percent similarity ($>70\%$) to known non-viral genomes were considered to be of potential viral origin and retained in the analysis. Of the contigs that were deemed to be of potential viral origin, those that were stable across all three body locations were considered to be additional putative viral markers for human identification, though their exact taxonomic classification is not known because they did not show sequence similarity or homology to any reference sequences.

In this study we identified a total of 188 viral markers, which included all stable NCBI reference mapped species, stable viral contig annotated species, and stable potential viral contigs that showed limited sequence similarity to databases or included viral-like gene regions. Of those 188 viral markers, 59 markers were found to be stable across all three body site locations sampled (Fig. 4D). These 59 target viral markers are proposed as having potential for the purpose of human identification, as shown across 42 individual test subjects. Of the 59 viral markers, seven were persistent across $>90\%$ of the sample population (Fig. 5). These seven markers may not be usable for subject-to-subject individualization, in regard to their presence or absence in the virome of an individual due to their low power of discrimination. However, they are prime candidates for genetic polymorphism analysis across a sampled population which we suspect will add an additional level of marker exploitation at the scale of single nucleotide variation, as well as presence/absence of insertions and deletions at the nucleotide level. As for the other 52 markers, they may be used as a presence/absence basis for human identification.

Distinct profile patterns were observed across the subjects (Fig. 5) for specific viral taxa. In particular, for subject participants P15-P21, they shared similar presence/absence profiles for certain viruses such as that of *Gammapapillomaviruses* which were not observed in the rest of the population. The presence of these distinct profile subpopulations is not due to sample processing or sequencing contamination due to the randomization of processing of samples and is believed to be a true biological finding. However, it is noted that for the subpopulation of P15-P21 these subjects frequently shared physical space and objects or were a cohabiting partner of one of the members of the aforementioned group. Not only does this subpopulation stand out from the rest of the population, but they can also be distinguished from one another based on their profiles. This suggests that not only can the virome be used as an individualizing characteristic but also bears circumstantial lifestyle or environmental characteristics. This data indicates that additional work is merited to examine subpopulations of people that cohabitate or share working environments.

The identified markers were found to be more significantly similar within subjects across time points compared to between subjects using presence/absence of the three identified marker sets. The more markers that were used (i.e., the combination of all sets) resulted in a greater significance in the similarity differences ($P = 5.3 \times 10^{-15}$). Thus, the addition of sets B and C were necessary in development of a more stable set for human identification profile production and evaluation. Diversity across subjects was also evaluated. We found that subject to subject variation was a significant variable associated with both α-diversity and β-diversity. Therefore, showing that not only are these viral markers stable but there is significant difference in subject-to-subject diversity which highlights the ability to separate the skin virome of one individual from another.

5. Conclusions

Viral biomarkers were identified from human skin virome metagenome samples from 42 individuals with the goal of developing and assessing the potential for human identification and diagnostics. In total, we identified 59 putative markers and we found seven markers that were present across all subjects and have potential to be used as targets for future studies into SNP and genetic variation within target viruses that could be used for discrimination of individuals within the population. We found the remaining 52 markers, when taken as a total community on the basis of presence and absence, were statistically significant across subjects ($P = 0.002$) and thus act as a full set of markers for human identification profile production.

Funding sources

This work was supported by the Department of Justice, USA [Grant numbers 2017-IJ-CX-0025 and 2019-75-CX-0017] and NIJ Fellowship [Grant number 2019-R2-CX-0048]. The funding agencies had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Declaration of Competing Interest

None to declare.

Data Availability

All raw sequencing data has been deposited in the NCBI Short Read Archive (SRA) under the accession code PRJNA754140. All metadata, list of markers, contig sequences, annotation files, and scripts described in the material and methods are publicly available and archived at: https://github.com/HerrLab/Graham_2021_forensics_human_virome.

Acknowledgments

We thank all participants for their contribution and participation in this study. We also thank W. Tom, A. Neujahr, N. Aluthge, C. Anderson, and others in the Fernando Lab who assisted with training of experimental methods and bioinformatic support. This work was completed using the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

Author contributions

Author contributions: MA, JC, SF, and JH contributed to the experimental design and conceptualization of the study; MA was responsible for participant recruitment and sample and metadata collection; EG and SF developed virome processing methodology and prepared samples for sequencing; EG and JH developed the bioinformatic pipeline and analyzed sequence data; JC contributed statistical support; EG, MA, JC, SF, and JH drafted the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102662.

References

[1] T.H. Clarke, A. Gomez, H. Singh, K.E. Nelson, L.M. Brinkac, Integrating the microbiome as a resource in the forensics toolkit, Forensic Sci. Int. Genet. 30 (2017) 141–147, https://doi.org/10.1016/j.fsigen.2017.06.008.

[2] U.A. Perego, A. Achilli, J.E. Ekins, L. Milani, M. Lari, E. Pilli, A. Brown, E.P. Price, S.R. Wolken, M. Matthews, C.A. Allen, T.R. Pearson, N. Angerhofer, D. Caramelli, T. Kupferschmid, P.S. Keim, S.R. Woodward, The Mountain Meadows Massacre and "poisoned springs": scientific testing of the more recent, anthrax theory, Int. J. Leg. Med. 127 (2013) 77–83, https://doi.org/10.1007/s00414-012-0681-y.

[3] C. Olivieri, I. Marota, F. Rollo, S. Luciani, Tracking plant, fungal, and bacterial DNA in honey specimens, J. Forensic Sci. 57 (2012) 222–227, https://doi.org/10.1111/j.1556-4029.2011.01964.x.

[4] S. Sathirareuangchai, P. Phuangphung, A. Leelaporn, V. Boon-yasidhi, The usefulness of Neisseria gonorrhoeae strain typing by pulse-field gel electrophoresis (PFGE) and DNA detection as the forensic evidence in child sexual abuse cases: a case series, Int. J. Leg. Med. 129 (2015) 153–157, https://doi.org/10.1007/s00414-014-1007-z.

[5] J.T. Hampton-Marcell, P. Larsen, T. Anton, L. Cralle, N. Sangwan, S. Lax, N. Gottel, M. Salas-Garcia, C. Young, G. Duncan, J.V. Lopez, J.A. Gilbert, Detecting personal microbiota signatures at artificial crime scenes, Forensic Sci. Int. 313 (2020), 110351, https://doi.org/10.1016/j.forsciint.2020.110351.

[6] H. Watanabe, I. Nakamura, S. Mizutani, Y. Kurokawa, H. Mori, K. Kurokawa, T. Yamada, Minor taxa in human skin microbiome contribute to the personal identification, PLoS One 13 (2018), https://doi.org/10.1371/journal.pone.0199947.

[7] S.E. Schmedes, A.E. Woerner, B. Budowle, Forensic human identification using skin microbiomes, Appl. Environ. Microbiol. 83 (2017), https://doi.org/10.1128/AEM.01672-17.

[8] H. Ikegaya, H. Iwase, C. Sugimoto, Y. Yogo, JC virus genotyping offers a new means of tracing the origins of unidentified cadavers, Int. J. Leg. Med. 116 (2002) 242–245, https://doi.org/10.1007/s00414-002-0297-8.

[9] M.R. Wilson, S.C. Weaver, R.A. Winegar, Legal, technical, and interpretational considerations in the forensic analysis of viruses, J. Forensic Sci. 58 (2013) 344–357, https://doi.org/10.1111/1556-4029.12065.

[10] S.E. Schmedes, A.E. Woerner, N.M.M. Novroski, F.R. Wendt, J.L. King, K. M. Stephens, B. Budowle, Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification, Forensic Sci. Int. Genet. 32 (2018) 50–61, https://doi.org/10.1016/j.fsigen.2017.10.004.

[11] H. Xu, H. Li, Acne, the skin microbiome, and antibiotic treatment, Am. J. Clin. Dermatol. 20 (2019) 335–344, https://doi.org/10.1007/s40257-018-00417-3.

[12] A. Reyes, M. Haynes, N. Hanson, F.E. Angly, A.C. Heath, F. Rohwer, J.I. Gordon, Viruses in the faecal microbiota of monozygotic twins and their mothers, Nature 466 (2010) 334–338, https://doi.org/10.1038/nature09199.

[13] S. Minot, R. Sinha, J. Chen, H. Li, S.A. Keilbaugh, G.D. Wu, J.D. Lewis, F. D. Bushman, The human gut virome: Inter-individual variation and dynamic response to diet, Genome Res. 21 (2011) 1616–1625, https://doi.org/10.1101/gr.122705.111.

[14] S. Minot, A. Bryson, C. Chehoud, G.D. Wu, J.D. Lewis, F.D. Bushman, Rapid evolution of the human gut virome, Proc. Natl. Acad. Sci. USA 110 (2013) 12450–12455, https://doi.org/10.1073/pnas.1300833110.

[15] G.D. Hannigan, J.S. Meisel, A.S. Tyldsley, Q. Zheng, B.P. Hodkinson, A. J. Sanmiguel, S. Minot, F.D. Bushman, E.A. Grice, The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome, mBio 6 (2015) e01578–15, https://doi.org/10.1128/mBio.01578-15.

[16] L.A. Ogilvie, B.V. Jones, The human gut virome: a multifaceted majority, Front. Microbiol. 6 (2015) 918, https://doi.org/10.3389/fmicb.2015.00918.

[17] N. Rascovan, R. Duraisamy, C. Desnues, Metagenomics and the human virome in asymptomatic individuals, Annu. Rev. Microbiol. 70 (2016) 125–141, https://doi.org/10.1146/annurev-micro-102215-095431.

[18] D. Sweet, M. Lorente, J.A. Lorente, A. Valenzuela, E. Villanueva, An improved method to recover saliva from human skin: the double swab technique, J. Forensic Sci. 42 (1997) 14120J, https://doi.org/10.1520/jfs14120j.

[19] B.C.M. Pang, B.K.K. Cheung, Double swab technique for collecting touched evidence, Leg. Med. 9 (2007) 181–184, https://doi.org/10.1016/j.legalmed.2006.12.003.

[20] S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data [Online], 2010. Available online at: ⟨http://www.bioinformatics.babraham.ac.uk/projects/fastqc/⟩.

[21] N. Joshi, J. Fass, Sickle: A Sliding-window, Adaptive, Quality-based Trimming tool for FastQ files (Version 1.33) [Software], 2011. Available at: ⟨https://Github.Com/Najoshi/Sickle⟩; ⟨https://github.com/najoshi/sickle⟩.

[22] B. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner, 2014. ⟨https://www.osti.gov/servlets/purl/1241166⟩.

[23] D. Li, C.M. Liu, R. Luo, K. Sadakane, T.W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics 31 (2015) 1674–1676, https://doi.org/10.1093/bioinformatics/btv033.

[24] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, https://doi.org/10.1093/bioinformatics/btt086.

[25] H. Li, Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM, 2013. ⟨http://arxiv.org/abs/1303.3997⟩.

[26] S. Nayfach, A.P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, N.C. Kyrpides, CheckV assesses the quality and completeness of metagenome-assembled viral genomes, Nat. Biotechnol. 39 (2021) 578–585, https://doi.org/10.1038/s41587-020-00774-7.

[27] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 76230, https://doi.org/10.1186/s13059-019-1891-0.

[28] Feargalr, Demovir: Taxonomic Classification of Viruses at Order and Family Level, 2019. ⟨https://github.com/feargalr/Demovir⟩.

[29] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.
[30] P. Menzel, K.L. Ng, A. Krogh, Fast and sensitive taxonomic classification for metagenomics with Kaiju, Nat. Commun. 7 (2016) 11257, https://doi.org/10.1038/ncomms11257.
[31] S.S. Langmead B, Fast gapped-read alignment with bowtie2, Nat. Methods 9 (2012) 357–359.
[32] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.
[33] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. ⟨https://www.R-project.org/⟩.
[34] P.J. McMurdie, S. Holmes, Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data, PLoS One 8 (2013), e61217, https://doi.org/10.1371/journal.pone.0061217.
[35] A.N. Shkoporov, A.G. Clooney, T.D.S. Sutton, F.J. Ryan, K.M. Daly, J.A. Nolan, S. A. McDonnell, E.V. Khokhlova, L.A. Draper, A. Forde, E. Guerin, V. Velayudhan, R. P. Ross, C. Hill, The human gut virome is highly diverse, stable, and individual specific, Cell Host Microbe 26 (2019) 527–541.e5, https://doi.org/10.1016/j.chom.2019.09.009.
[36] D.P. Faith, P.R. Minchin, L. Belbin, Compositional dissimilarity as a robust measure of ecological distance, Vegetatio 69 (1987) 57–68, https://doi.org/10.1007/BF00038687.
[37] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
[38] J. Oksanen, F. Guillaume Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M. Henry, H. Stevens, E. Szoecs, H. Wagner, Vegan: Community Ecology Package, 2020. ⟨https://cran.r-project.org/package=vegan⟩.
[39] J. Oh, A.L. Byrd, M. Park, H.H. Kong, J.A. Segre, Temporal stability of the human skin microbiome, Cell 165 (2016) 854–866, https://doi.org/10.1016/j.cell.2016.04.008.
[40] S. Roux, M. Krupovic, D. Debroas, P. Forterre, F. Enault, Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences, Open Biol. 3 (2013), 130160, https://doi.org/10.1098/rsob.130160.
[41] G. Liang, F.D. Bushman, The human virome: assembly, composition and host interactions, Nat. Rev. Microbiol. 19 (2021) 514–527, https://doi.org/10.1038/s41579-021-00536-5.
[42] O. Tirosh, S. Conlan, C. Deming, S.Q. Lee-Lin, X. Huang, B.B. Barnabas, G. G. Bouffard, S.Y. Brooks, H. Marfani, L. Dekhtyar, X. Guan, J. Han, S. ling Ho, R. Legaspi, Q.L. Maduro, C.A. Masiello, J.C. McDowell, C. Montemayor, J. C. Mullikin, M. Park, N.L. Riebow, K. Schandler, C. Scharer, B. Schmidt, C. Sison, S. Stantripop, J.W. Thomas, P.J. Thomas, M. Vemulapalli, A.C. Young, H.C. Su, A. F. Freeman, J.A. Segre, H.H. Kong, Expanded skin virome in DOCK8-deficient patients, Nat. Med. (2018) 1815–1821, https://doi.org/10.1038/s41591-018-0211-7.
[43] G. Trubl, H. Bin Jang, S. Roux, J.B. Emerson, N. Solonenko, D.R. Vik, L. Solden, J. Ellenbogen, A.T. Runyon, B. Bolduc, B.J. Woodcroft, S.R. Saleska, G.W. Tyson, K. C. Wrighton, M.B. Sullivan, V.I. Rich, Soil viruses are underexplored players in ecosystem carbon processing, mSystems 3 (2018), https://doi.org/10.1128/mSystems.00076-18.
[44] A.C. Gregory, O. Zablocki, A.A. Zayed, A. Howell, B. Bolduc, M.B. Sullivan, The gut virome database reveals age-dependent patterns of virome diversity in the human gut, Cell Host Microbe 28 (2020) 724–740.e8, https://doi.org/10.1016/j.chom.2020.08.003.
[45] V. Foulongne, V. Sauvage, C. Hebert, O. Dereure, J. Cheval, M.A. Gouilh, K. Pariente, M. Segondy, A. Burguière, J.C. Manuguerra, V. Caro, M. Eloit, Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing, PLoS One 7 (2012), e38499, https://doi.org/10.1371/journal.pone.0038499.
[46] L.J. van Zyl, Y. Abrahams, E.A. Stander, B. Kirby-McCollough, R. Jourdain, C. Clavaud, L. Breton, M. Trindade, Novel phages of healthy skin metaviromes from South Africa, Sci. Rep. 8 (2018) 12265, https://doi.org/10.1038/s41598-018-30705-1.
[47] D.V. Pastrana, A. Peretti, N.L. Welch, C. Borgogna, C. Olivero, R. Badolato, L. D. Notarangelo, M. Gariglio, P.C. FitzGerald, C.E. McIntosh, J. Reeves, G. J. Starrett, V. Bliskovsky, D. Velez, I. Brownell, R. Yarchoan, K.M. Wyvill, T. S. Uldrick, F. Maldarelli, A. Lisco, I. Sereti, C.M. Gonzalez, E.J. Androphy, A. A. McBride, K. Van Doorslaer, F. Garcia, I. Dvoretzky, J.S. Liu, J. Han, P. M. Murphy, D.H. McDermott, C.B. Buck, Metagenomic discovery of 83 new human papillomavirus types in patients with immunodeficiency, mSphere 3 (2018) e00645–18, https://doi.org/10.1128/mspheredirect.00645-18.
[48] M.J. Tisza, D.V. Pastrana, N.L. Welch, B. Stewart, A. Peretti, G.J. Starrett, Y.Y. S. Pang, S.R. Krishnamurthy, P.A. Pesavento, D.H. McDermott, P.M. Murphy, J. L. Whited, B. Miller, J. Brenchley, S.P. Rosshart, B. Rehermann, J. Doorbar, B. A. Ta'ala, O. Pletnikova, J.C. Troncoso, S.M. Resnick, B. Bolduc, M.B. Sullivan, A. Varsani, A.M. Segall, C.B. Buck, Discovery of several thousand highly diverse circular DNA viruses, elife 9 (2020), e51971, https://doi.org/10.7554/eLife.51971.
[49] A.L. Byrd, Y. Belkaid, J.A. Segre, The human skin microbiome, Nat. Rev. Microbiol. 16 (2018) 143–155, https://doi.org/10.1038/nrmicro.2017.157.
[50] N.-D. van Loo, E. Fortunati, E. Ehlert, M. Rabelink, F. Grosveld, B.J. Scholte, Baculovirus infection of nondividing mammalian cells: mechanisms of entry and nuclear transport of capsids, J. Virol. 75 (2001) 961–970, https://doi.org/10.1128/jvi.75.2.961-970.2001.
[51] J.D. Siqueira, G. Curty, D. Xutao, C.B. Hofer, E.S. Machado, H.N. Seuánez, M. A. Soares, E. Delwart, E.A. Soares, Composite analysis of the virome and bacteriome of HIV/HPV co-infected women reveals proxies for immunodeficiency, Viruses 11 (2019) 422, https://doi.org/10.3390/v11050422.
[52] E.M. Bolatti, T.M. Zorec, M.E. Montani, L. Hošnjak, D. Chouhy, G. Viarengo, P. E. Casal, R.M. Barquez, M. Poljak, A.A. Giri, A preliminary study of the virome of the south american free-tailed bats (Tadarida brasiliensis) and identification of two novel mammalian viruses, Viruses 12 (2020), https://doi.org/10.3390/v12040422.