Document Title:	The Human Virome as Trace Evidence in Forensic Investigation
Author(s):	Michael Adamowicz, Ph.D.
Document Number:	304614
Date Received:	April 2022
Award Number:	2017-IJ-CX-0025

Final Report

DOJ Grant 2017-IJ-CX-0025

Project Titled; The Human Virome as Trace Evidence in Forensic Investigation

Michael Adamowicz, Ph.D.

madamowicz2@unl.edu

103 Agriculture Hall

Lincoln, NE 68583

402-472-7932

Board of Regents, University of Nebraska-Lincoln

151 Whittier Research Center

2200 Vine Street

Lincoln, NE 68583

Grant Period 01/01/2018 – 12/31/2020

Award $698,382

Summary of project

This project was designed to address the ongoing need to create forensically relevant linkages between persons, places, and objects by developing the untapped potential of the human viral microbiome (*virome*). The human virome is a source of rich genetic diversity that needs to be examined to determine if it is stable, transferable, and provides sufficient power of discrimination to be used as an alternative to traditional human forensic deoxyribonucleic acid (DNA) tests when conditions for such tests are not suitable. The human bacterial microbiome has already been examined as an alternative method for post-mortem interval determination and as a marker in cases involving soil samples. The human virome offers some advantages, as viral genomes are even smaller than those of bacteria, and thus are potentially more stable; have a variety of morphologies (double- and single-stranded), increasing the possible number of discriminating markers; and is present throughout the human body, including the skin and bodily fluids, making it transferrable. Also, the copy number of viral genomes in a given volume is substantially higher, compared to the copy number of human or bacterial genomes, increasing the likelihood of isolating a sufficient quantity for successful testing. There has been a complete lack of empirical data on the human virome regarding its suitability to forensic applications. The work described in this report has begun to address this gap in our knowledge by generating data in the previously mentioned areas of stability, transfer, and discrimination. The proposal hypothesized that the genetic diversity contained in each human being's particular skin virome could be translated into a pattern profile that could be used to discriminate them from other people and that the virome profile could be detected when standard human DNA samples were not viable. The work has been completed with protocols, instrumentation, and analysis methods that are either already in forensic DNA laboratories or can be readily assimilated. This work has led to increasing our understanding of an additional, powerful comparative DNA tool for cases in which human DNA is too

scant or degraded to support the derivation of a statistically useful profile, thus adding individualizing information where there would have been none.

The objectives of the research project have been to develop new technology as follows: 1) create a two-pronged discriminatory approach to generating human virome profiles that leveraged the advantages of reference dependent and independent methods i.e. denovo and reference-based analyses 2) assess the forensic significance of the profiles in terms of their efficacy in separating individual humans, taking cohabitation of individuals into account and 3) create a web-accessible public database of human skin virome profiles for the estimation of composition and rarity of such profiles. The projects prime questions were whether sufficient viral particles could be harvested from a person's skin to yield a virome profile and would that profile be stable over time such that it could provide a reliable level of discrimination between people in a defined population.

Skin swab samples were collected, using a dry and wet swab. The wet swab was moistened with sterilized 1x phosphate buffered saline (PBS). The swabs were used together for collecting virome samples. Virome samples were collected from 42 individuals with ages ranging from 19 – 70+ years. Samples were collected across a six-month period to represent day 0 (initial swab), 2-weeks, 1-month, 3-month, and 6-month from the initial sampling. At each collection, virome swabs were collected from three skin locations (left hand, right hand, and scalp). At each sampling the participants completed a questionnaire to gather information on travel, grooming, lifestyle, and other information that could help identify features affecting the microbiome. During each sampling, negative control swabs were collected to evaluate any contamination that may result from environmental factors including the PBS used. The swabs collected were stored at -20°C until used for viral enrichment and sequencing.

Swabs containing skin viromes were saturated with 200µl of 0.02µm filtered sterile 1x PBS and were placed in a 2ml tube containing a spin basket. Swabs were centrifuged at 16000 x g for 10 minutes to elute viral particles from the swab into the PBS solution. The filtrate containing the viral particles was further filtered using a 0.22µm filter to remove cellular and bacterial contaminants. The resulting filtrate was enriched for viral particles and was used for viral DNA extraction using the QIAamp Ultra-Sensitive Virus Kit according to the manufacturer's protocol. The resulting viral DNA was subjected to whole genome amplification (WGA) using multiple displacement amplification (MDA) using the TruePrime WGA Kit (Syngin) and associated protocol. Following WGA, the samples were quantified using the DeNovix dsDNA High Sensitivity Kit.

One hundred nanograms of the amplified viral DNA was used for library preparation. The DNA was sheared using sonication to 600bp in length. The resulting sheared DNA was used for library preparation using the NEBNext Ultra II Library preparation kit (New England Biolab) according to the manufacturer's protocol. The DNA fragments were then blunt ended and phosphorylated. Size selection of 500-700bp fragments was performed with magnetic bead-based purification. Final library preps were evaluated using an Agilent Bioanalyzer with high sensitivity chips to identify sample base pair distribution and sample concentration. Additionally, libraries were quantified using the DeNovix dsDNA High Sensitivity Kit. Resulting libraries were sequenced using the 150 bp paired-end sequencing strategy on the Illumina Hiseq 2500 platform.

The subsequent sequencing data was trimmed and filtered using the adaptive trimming tool Sickle v.1.3.3 to remove low quality reads using a quality filter threshold of Q30 and a length threshold of 75 bp. Reads resulting from Phi X were removed from trimmed reads using the BBDuk. Following quality filtering, bacterial contamination was assessed by mapping trimmed reads to the Silva 16S ribosomal database v.138.1 with BBMap using parameters described for high precision mapping of contamination detection. Additionally, all sample reads were mapped to the human genome (hg19) using BBMap as per BBMap Guide, for high precision mapping with low sensitivity to lower the risk of false positive mapping. All mapped reads were removed to ensure the viromes are devoid of contamination from bacterial and human host. Metagenome assemblies were performed using Megahit v.1.2.8. 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. In addition, a meta-assembly using all negative control reads was performed to identify potential contaminants in the dataset that may have arisen from reagents. The virome assemblies generated were mapped to the negative control contigs greater than 1000bp using BWA-mem to remove any reads that may have resulted from contamination. Subsequent contigs greater than 1000bp were then utilized for downstream analysis for viral identification, diversity analysis, and assessment of stability of the virome.

Putative viral contigs containing viral genes were identified using the tool CheckV v.0.7.0. Contigs that were determined to have viral genes, as per CheckV results, were identified as viral sequences and were classified using various viral annotation and viral classification tools as described below. Viral contigs were classified using both nucleotide-based classification tools Kraken2 v.2.0.8-beta, Demovir, Blastn (with a >10% query coverage cut-off) and using a protein-based classification tool Kaiju v.1.7. After assessing the results from all the classification tools, the resulting classification having the lowest e-value or highest percent confidence was used. The least common ancestor was used for classification results having similar e-values or percent confidences but different results to reduce misclassification of viral contigs.

Raw sample reads were mapped to the meta-assembly consisting of contigs from the sample-by-sample based assembly. Read mapping was performed using Bowtie 2 v.2.3.5 and subsequent Samtools v.1.9 manipulation to identify the abundance of each viral gene/contig within each sample. The resulting viral gene abundances were further analyzed using “R” v.3.6.3. Unique contigs were used to identify viral diversity and changes over-time. A phyloseq object was created using viral/gene abundance data and was used for diversity analysis. Annotation of viral contigs was performed using the classification tools described above. The count table and mapping file information was used as input for phyloseq object generation, allowing for both denovo and reference-based analysis of the skin virome data.

Viral Contig stability was assessed based on presence and absence of viral contigs over time on each body site sampled, left hand, right hand, or scalp, within each subject. Contigs present in 4 out of the 5 time points from a location within an individual was considered to be a stable viral contig and was identified as a potential marker for human identification. Out of the identified viral contigs, those contigs that were not identified in both hands for a given individual were removed. Viral gene contig stability was assessed at family, genus, and species level. Additionally, an assembly independent method was employed using the identified stable viral families for further refinement of viral taxonomic identification at the species and genus level and investigation into punitive human identification makers. Briefly, all sequence files greater than 1000bp were identified from the NCBI nucleotide database classified to families Papillomaviridae, Genomoviridae, Baculoviridae, and the order Caudovirales and were downloaded for subsequent analysis of mapping reads to their families. Raw reads were mapped to the reference sequences identified as belonging to families Papillomaviridae, Genomoviridae, Baculoviridae, and the order Caudovirales using Bowtie 2. The resulting mapped reads were further analyzed using Samtools to acquire read counts. The read counts generated were utilized for further investigation of viral diversity and persistence of selected viral families and for statistical evaluation using R.

Alpha diversity was assessed using the Shannon and inverted Simpsons alpha diversity metrics. To evaluate if a subject's contig diversity significantly changed across body sites and time, a one-way ANOVA using repeated measures was used. For Beta diversity, a Bray-Curtis dissimilarity matrix was generated based on contig distribution and read abundance and was used for PERMANOVA analysis to assess changes of the virome between subjects. The data was visualized using a principal coordinate analysis (PCoA).

Gathering data on these objectives has furthered our understanding of the human viral microbiome, addressing the critical barrier of individual identification when conditions for human DNA testing are poor. The project has contributed to developing new technology for the use of the human virome as a tool for forensic applications. The advantages of opening another area of forensic biological testing are significant to the criminal justice system. Having the potential to generate a type of DNA profile when the normal methods for human identification are not working could be the only individualizing evidence in a case. There are some types of cases that do not lend themselves to the current human DNA testing methods, as the nature of the biological material transferred is problematic, such as digital-penetration sexual assault cases. These can produce mixed DNA profiles wherein the suspect's DNA is so dilute and at such a low concentration, that no profile can be derived. The suspect's virome, however, may be detectable and discernable from that of the victim based on both individual genetic and general *genus* markers, similar to work that has been done with Neisseria gonorrhoeae, but on a finer level. There are also many cases that yield only partial human DNA profiles, stochastic mixed profiles, or a combination of the two. These can be very difficult to interpret, often requiring advanced probabilistic genotyping software to gain any interpretive value. The addition of human virome markers could complement the existing data and provide added value, particularly in making a clear exclusion in an otherwise inconclusive interpretation.

Participants

Name: Jennifer Clarke

Project Role: Co-PI

Name: Samodha Fernando

Project Role: Co-PI

Name: Joshua Herr

Project Role: Co-PI

Name: Ema Graham

Project Role: Graduate Student (Ph.D. candidate)

Changes in approach from original design

There were no changes in the project from the original design. Analysis of the data collected will go on for some time, as the dataset is enormous and further discovery is expected. The last 18 participant samples are still in the process of sample sequencing and analysis and data from them is not included in the report, however sample collection for all proposed collections was completed.

Outcomes

Skin virome samples were collected across three skin site locations (left hand, right hand, and scalp) from 60 individuals, with 42 being completed through full data analysis. Differing skin locations were used in this study to evaluate differing skin environments. These locations were chosen due to their potential to be of forensic relevance and their higher levels of viral abundance as shown in previous studies. Both hands were swabbed at time of collection to act as a replicate and to evaluate differences in virome composition and diversity across similar skin types but having differing environmental contact due to things such as hand dominance. In addition to evaluation of differing skin site locations virome stability over time was also assessed. A 6-month time course was used to assess seasonal viral diversity and limit seasonal effects in forensic viral marker determination.

Once collected samples were extracted, whole genome amplified, and sequenced the resulting data was bioinformatically processed to remove low quality reads and contamination. Bacterial contamination was assessed by mapping trimmed reads to 16S reference reads obtained from the Silva database v.1.3.7. Percentages of trimmed reads per sample mapped to 16S reads ranged from 0% to 0.15%, showing lowered levels of bacterial contamination in samples and thus showing sufficient virome sequencing preparation sample processing. Trimmed and processed reads were then assembled further processed by removing negative control mapped contigs from the overall assembly. All contigs 1000bp in length and larger were retained resulting in 62,101 contigs. Contigs were run through CheckV and of the 62,101 assembled contigs, 1298 were identified as having a known viral gene and considered to be viral in origin. These constitute the current working dataset.

Detected viral taxonomy was similar to that seen in previous studies. Of the annotated viral contigs, the majority of the contigs, though containing a viral gene and being identified as being of viral origin, were unable to be classified and had low percentage identity hits to any known sequences in NCBI blast demonstrating the highly limited current state of viral database availability and the vast unexplored nature of viral diversity and taxonomy. Though unclassified viruses were highly abundant across samples, both double stranded DNA viruses and single stranded viruses were also abundant and able to be detected and annotated in the project’s dataset. Of the double stranded DNA viruses, the Order of Caudovirales was the most abundantly detected. This is not surprising, seeing as how there is a disproportionate amount of Caudovirales sequenced genomes available in viral reference databases due to their ability to be cultured. Viral families observed in the skin virome that fall under the Order of Caudovirales, such as that of Siphoviridae, Podoviridae, and Herelleviridae, are bacteriophage that are associated with infection of bacteria that are commonly associated with the skin microbiome such as Staphylococcus and Streptococcus. As for the identified single stranded viruses, the most abundant taxa were those of the small circular DNA viruses such as papilloma viruses and Cress-like DNA pages. Papilloma and polyoma viruses are common skin associated opportunistic pathogens and the identification of papilloma viral genomes was expected. However, the identification of small cress-like DNA phages has not been reported in previous studies. This is due to the recent discovery and annotation of novel Cress-DNA viruses and their addition to NCBI databases. It is also important to note that there were many unclassified viruses that were small and circular in nature and had similarity to that of small circular DNA viruses in viral reference databases. However, due to their low percent identity to any known virus they were unable to be classified down to either a species or genus level. This was especially observed in unclassified viruses having similarity to Microviridae classified viruses. Thus, the skin virome contains many novel viruses and this work demonstrates more study into viral discovery is needed for the full picture of viral diversity to be assessed.

The top ten most abundant viral families were identified for each location collected on a subject-to-subject basis. Of the most abundant viral families observed per person, those of note are the Papillomaviridea and varying viral families that fall under the order of Caudovirales. This is consistent with the findings of the overall virome diversity assessment across all individuals. Similarities across all three locations for each individual were noted. Of the locations, the scalp and right hand shared similar relative abundance of these viral families. Left hand, though similar, did show some variation from that of the right and the scalp as seen for subjects P05, P15, P28, P32, P40, P42, and P43. For these instances there was either a complete loss of a highly abundant family as compared to that of the right hand or the scalp or the addition of a viral family such as that of Uncultured virus and Streptococcus satellite phage Javan 305. However due to the incomplete annotation of these two mentioned viral families, they may fall into a different viral family that is already being accounted for such as that of Siphoviridae. Due to the incomplete status of viral database information, further investigation it is still needed. Therefore, the outcomes of this study recommend that viral target marker identification should be performed at an evaluation of a genus or species level to alleviate database annotation inaccuracies and naivety. As for a comparison of person-to-person relative abundance of viral families, there are clear differences across individuals. Just using the most abundant viral families, clear distinctions can be drawn between individuals, while still maintaining higher levels of similarity within an individual across multiple skin collection locations, further supporting the notion that the human skin virome is individualized and potentially be used for discrimination between differing subjects.

Evaluation of the viromes most abundant viral family level taxa shows higher levels of difference between individuals while maintaining similarities within an individual. Though highly abundant, for viral markers to be sufficient for forensic evidentiary material, these markers must also be stable over time thus allowing for comparison of collected and generated viral profiles collected at different time points. Therefore, prevalence of contig annotated viral families was evaluated for stability across individuals and locations within an individual. Stable viral taxa were categorized by presence in at least 4 out of the 5 time points for an individual at a certain skin location. The viral annotated families of Baculoviridae, Genomoviradea, Herelleviridea, Myoviridae, Papillomaviridae, Podoviridae, Siphoviridae, Unclassified Caudovirales, and Unclassified Homo sapiens like virus were stable in at least one individual across all three skin locations and thus considered potential target viral families for further investigation for utilization as human identification target markers. Though these viral families may have stability in certain individuals, there are higher incidences of temporal precedence in virome samples overall. Thus, viral families cannot alone be utilized as a taxonomic marker for human identification and a higher level of annotation must be used such as that of genus or species.

To improve viral genome recovery and reduce metagenome assembly bias, trimmed and contamination removed reads were mapped to all NCBI nucleotide viral reference sequences associated with target viral families identified to be stable across all three skin locations within at least one individual from the annotated viral contigs. All NCBI nucleotide viral sequences associated with the Order of Caudovirales were used due to the fact that multiple target families, such as Siphoviridae, Podoviridae, and Myoviridae, all fall under the Order Caudovirales. The stability of these mapped counts to the reference genomes was evaluated the same way as was done prior but on a species level as opposed to at a family level evaluation. In addition, annotated contig stability was also evaluated for species level marker identification. Viral species found to be stable in 4 out of the 5 time points within a location for at least one individual were considered stable viral species.

To address viral dark matter that is unable to be annotated and viral identification bias based on only known viral genes, all contigs were also subjected to evaluation for stability as mapped viral reference species and annotated viral contig species. NCBI's BlastN was used to evaluate contig sequences that were considered to be stable for genome similarity to known organisms. Contigs containing <70% identity to a known organism or having the presence of both prokaryotic and eukaryotic genes with regions of 0% contig coverage between adjacent to identified genes were considered to be putative viral sequences though viral origin cannot be fully confirmed.

In total, of the stable viral species and dark matter contigs, 197 were determined to be stable and thus proposed as potential target makers. To further limit the number of markers, only markers that were found to be stable across all three locations within at least one individual were retained for marker determination, resulting in 62 putative human identification target markers. Of the target markers, 7 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 discriminatory power across individuals.

Assembled viral contigs were identified using viral gene identification thus generating a dataset of contigs of known viral origin. Of the 62,101 assembled contigs, 1298 were able to be identified as viral. This is a small percentage of the overall metagenome assembly. However, this small dataset is a conservative consortium of contigs and is an assumed underestimation of the true number of viral contigs. Since viral contigs were determined by using bioinformatic tools that identify known viral genes, fractured smaller contigs that do not contain full gene sequences or contigs of novel viruses that display enough genetic differentiation from that of known viral genes in the databases used by the tools will escape viral identification and thus were not included in the viral contig dataset. Due to regions of repetition, retention of genes from host genomes, high levels of nucleotide sequence diversity by mutations, and genetic motifs such as inverted terminal repeats commonly seen in viruses, viral metagenomic sequence data can lead to erroneous assembly and assemblies with increased fractionation than that of non-viral organism related sequencing data. The results of this study should be viewed as a proof of concept that the human virome offers a rich source of biomarkers that can be used for human identification. A substantial number of viral markers have been identified and characterized as stable, human skin markers. The target marker information is being used in the follow up study to assess how well the viral particles can be transferred by touch or from telogenic hair samples. Further studies using PCR amplification and amplicon sequencing of conserved genes within target viral family and species level taxonomy are required for complete capture of viral sequence diversity and taxonomic presence.

The primary limitation in this project has been the lack of viral taxonomic data. This was expected from the beginning as the field of virology simply has no real grasp on just how diverse the world of viruses is. The two-pronged approach used in the study was designed to address this limitation and has been successful. We have identified over 60 stable human skin viral markers, to this point, that could be used for forensic human identification. This dataset has been accomplished using information from a large number of sample collections from a relatively small population (42 individuals). A larger population sample size will certainly yield even more viral markers for further study.

Artifacts

Conference Papers:

M.S. Adamowicz, J. Clarke, S. Fernando, E. H. Graham, J. Herr, and G. Watkins, 2021. Human Identification Using the Skin Virome. Presented at the AAFS 73rd Annual Scientific Meeting. Presentation B122.

M.S. Adamowicz, J. Clarke, S. Fernando, E. H. Graham, J. Herr, and G. Watkins, 2021. Human Identification Using the Skin Virome. Presented at Pittcon Conference & Expo, 03/11/21 Session G01-08.

Dataset:

To date 62,101 assembled contigs, with 1298 identified as viral and 62 putative human identification target markers. The dataset is being placed into National Center for Biotechnology Information's short read archive.