Evolutionary and functional implications of hypervariable loci within the skin virome

Geoffrey D. Hannigan[1], Qi Zheng[1], Jacquelyn S. Meisel[1], Samuel S. Minot[2], Frederick D. Bushman[3] and Elizabeth A. Grice[1,3]

[1] Department of Dermatology, University of Pennsylvania, Philadelphia, PA, USA
[2] One Codex, San Francisco, CA, USA
[3] Department of Microbiology, University of Pennsylvania, Philadelphia, PA, USA

ABSTRACT

Localized genomic variability is crucial for the ongoing conflicts between infectious microbes and their hosts. An understanding of evolutionary and adaptive patterns associated with genomic variability will help guide development of vaccines and antimicrobial agents. While most analyses of the human microbiome have focused on taxonomic classification and gene annotation, we investigated genomic variation of skin-associated viral communities. We evaluated patterns of viral genomic variation across 16 healthy human volunteers. Human papillomavirus (HPV) and *Staphylococcus* phages contained 106 and 465 regions of diversification, or hypervariable loci, respectively. *Propionibacterium* phage genomes were minimally divergent and contained no hypervariable loci. Genes containing hypervariable loci were involved in functions including host tropism and immune evasion. HPV and *Staphylococcus* phage hypervariable loci were associated with purifying selection. Amino acid substitution patterns were virus dependent, as were predictions of their phenotypic effects. We identified diversity generating retroelements as one likely mechanism driving hypervariability. We validated these findings in an independently collected skin metagenomic sequence dataset, suggesting that these features of skin virome genomic variability are widespread. Our results highlight the genomic variation landscape of the skin virome and provide a foundation for better understanding community viral evolution and the functional implications of genomic diversification of skin viruses.

Subjects Microbiology, Dermatology
Keywords Bacteriophage, Genomic variability, Metagenomics, Evolution, Virome, Dermatology

INTRODUCTION

Localized genomic modifications are ammunition in the ongoing battle between hosts and infectious agents. The human adaptive immune response relies on localized genomic diversity of antigen receptors to facilitate detection and efficient removal of foreign agents (*Borghans, Beltman & De Boer, 2004*; *Kubinak et al., 2012*). Infectious microbes, such as bacteria and viruses, likewise rely on genomic variation to modulate tropism and facilitate immune evasion (*Malim & Emerman, 2001*; *Doulatov et al., 2004*; *Minot et al., 2012*; *Schillinger et al., 2012*; *Das et al., 2013*; *Minot et al., 2013*; *Guo et al., 2014*).

Submitted 7 October 2016
Accepted 5 January 2017
Published 7 February 2017

Corresponding author
Elizabeth A. Grice,
egrice@upenn.edu

Academic editor
Katrine Whiteson

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj.2959**

Potential selective benefits of targeted variation in viruses include immune evasion and widening of host tropism (*Borghans, Beltman & De Boer, 2004*; *Kubinak et al., 2012*).

Most contemporary low resolution studies of the human microbiome evaluate functional potential through taxonomic classification and whole gene identification (*Schloss & Handelsman, 2008*; *Human Microbiome Project Consortium, 2012*; *Langille et al., 2013*; *Hannigan et al., 2014*; *Ly et al., 2014*; *Norman et al., 2015*; *Lim et al., 2015*; *Meisel et al., 2016*). These approaches are usually unable to capture nucleotide variations that affect functionality of proteins encoded in the microbiome, which can be altered by differences in only a few nucleotides. For example, viruses such as *Bordetella* bacteriophages, hepatitis C virus, and others only require short variable regions within a gene to facilitate functional changes in processes including tropism diversity, immune evasion, drug resistance, and adaptation to host auxotrophies (*Bacher, Bull & Ellington, 2003*; *Doulatov et al., 2004*; *Donaldson et al., 2010*; *Guan et al., 2012*; *Shah et al., 2014*). Contemporary low-resolution studies also fail to identify genetic cassettes that promote targeted diversity, such as diversity generating retroelements (DGRs). DGRs promote targeted genetic diversification in bacteriophages through error-prone cycles of transcription, reverse transcription, and integration; through this process information encoded in a non-variable template region is copied in a fallible fashion into a variable region within a coding sequence (*Doulatov et al., 2004*; *Minot et al., 2012*; *Schillinger et al., 2012*).

Here, we investigate skin virome evolution and adaptation by inferring the selective pressure, functional diversity, and substitution patterns associated with targeted hypervariation. We focus on three prominent cutaneous viruses: human papillomavirus (HPV), *Propionibacterium* phage, and *Staphylococcus* phage. HPV is associated with the development of skin cancer, especially in immune-suppressed individuals (*Vinzón et al., 2014*; *Wang et al., 2014*; *Quint et al., 2015*). Current vaccine efforts aim to target conserved antigens for broad strain protection—thus a greater understanding of HPV genomic diversity could improve design of vaccines (*Schiller & Lowy, 2012*; *Vinzón et al., 2014*). *Staphylococcus* phages can modulate *Staphylococcus* pathogenic gene expression and facilitate transmission of antibiotic resistance (*Bae et al., 2006*; *Varga et al., 2012*). *Propionibacterium* phages are associated with *Propionibacterium acnes* and have therapeutic potential for treating acne (*Marinelli et al., 2012*; *Hannigan & Grice, 2013*). Our findings build upon previous analyses of individual virus genomic variability and provide new insight into phage biology of the cutaneous microbiome.

MATERIALS AND METHODS

Analysis details and availability

All associated source code and explanatory README files are available for review at the following GitHub repository: https://github.com/Microbiology/ViromeVarScripts.

Data acquisition and quality control

The primary skin virome dataset was acquired from SRA accession: SRP049645 (*Hannigan et al., 2015*) and includes sequences from samples collected at the second and

third time points. The secondary dataset was obtained from *Oh et al. (2014)* (SRA BioProject: 46333). Retroauricular crease samples were downloaded from the NCBI SRA BioProject: 46333. For samples from both the primary and secondary dataset, sequences were trimmed with the FASTX-Toolkit (version 0.0.14), using a quality score cutoff of 33. Remaining sequences with similarity to the human genome were removed using the standalone DeconSeq toolkit (version 0.4.3) (*Schmieder & Edwards, 2011*).

Contig assembly and taxonomic identification

Contigs from the primary dataset were obtained from the published Figshare source (DOI 10.6084/m9.figshare.1281248). Contigs from both the primary and secondary datasets were separately assembled using the Ray metagenomic assembly software, specifying a minimum contig length of 500 bp and otherwise default parameters (v2.3.1) (*Boisvert et al., 2012*). Within each dataset, sequences from all samples were combined prior to assembly to facilitate the most complete contig assembly. Contig coverage was determined by aligning sequences back to the contigs with the bowtie2 toolkit (v2.1.0; seed substring length of 25 and one mismatch allowed in alignment) (*Langmead & Salzberg, 2012*). Quantification of reads mapping back to contigs was obtained by parsing bowtie2 output using Perl and BASH scripts presented in the supplemental source code. Coverage was calculated using the number of reads mapping to each contig. The blastn program from the NCBI Blast+ toolkit (version 2.2.0) was used to determine similarity of contigs to virus reference genomes (*Camacho et al., 2009*). Contigs were blasted against a previously described virus-specific genome reference database, which is a subset of the EMBL reference genome database (*UniProt Consortium, 2014*; *Hannigan et al., 2015*). A similarity threshold of *e*-value $< 10^{-3}$ was used, and sequences with multiple potential identities were resolved by using only hits with the lowest *e*-values. Although this was the minimum threshold, the contigs of interest exhibited *e*-values $< 10^{-3}$.

Phylogenetic analysis

We constructed phylogenies using the L1 capsid gene for HPV (*Ma et al., 2014*) and the large terminase subunit for the *Staphylococcus* and *Propionibacterium* phages (*Gutiérrez et al., 2013*; *Ma et al., 2014*) as phylogenetic marker genes. For reference, we used the PAVE reference L1 genes (https://pave.niaid.nih.gov/, accessed 2015-06-03) (*Van Doorslaer et al., 2013*). The large terminase subunit references for *Staphylococcus* and *Propionibacterium* phages were from the NCBI gene sequence database (*Staphylococcus* phage: accessed 2015-09-14, search term: *((phage terminase large subunit staphylococcus)) AND "viruses"[porgn:__txid10239] NOT "ORF" NOT "hypothetical" NOT "putative;"* *Propionibacterium* phage: accessed 2015-09-15, search term: *((phage terminase large subunit propionibacterium)) AND "viruses"[porgn:__txid10239]*). To extract the phylogenetic marker genes from the virome contigs, we determined which open reading frames (ORFs) matched the reference genes by nucleotide similarity (nucleotide blast, *e*-value 1e−10). Only ORFs longer than 1.2 kb were included in the analysis. The average reference gene lengths were all longer than this threshold (average reference gene length: HPV = 2,519 bp, *Staphylococcus* phage = 1,307 bp, *Propionibacterium* phage = 1,511 bp).

Contig and reference marker genes were aligned using the Smith–Waterman algorithm and 1,000 iterations as implemented by the mafft aligner (v7.221) (*Katoh & Standley, 2013*). Phylogeny was constructed using RAxML (version 8.1.21) (*Stamatakis, 2014*). The phylogenetic tree was visualized using Figtree (*Rambaut, 2006*).

Identification of temperate phage contigs

As has been described previously, we identified temperate (lysogenic) phage contigs using three genomic markers: contig nucleotide similarity to (1) phage integrase genes, (2) prophage genes within the ACLAME prophage database, and (3) bacterial reference genomes. We performed a blastx alignment (*e*-value 1e−10, percent identity threshold 75%) of the genes within the ACLAME prophage database (*Leplae, Lima-Mendez & Toussaint, 2010*), a blastx alignment with integrase genes from Uniprot database, and a blastn alignment of the *Staphylococcus* phage contigs to *Staphylococcus* bacterial reference genomes. Integrase genes were obtained from the online Uniprot database using the search term "*organism:phage gene:int NOT putative*." *Staphylococcus* reference genomes were obtained from the NCBI nucleotide database using the search term "'Staphylococcus'[Organism] AND 'complete genome'[Name] NOT phage[All Fields] NOT contig[All Fields] NOT ('unidentified plasmid'[Organism] OR plasmid[All Fields]) AND (bacteria[filter] AND biomol_genomic[PROP])." Both were accessed December 22, 2016. Together this allowed us to detect regions of contigs that demonstrated a high similarity to temperate phage gene signatures.

Identification of hypervariable loci

The bowtie2 alignments of reads to viral contigs were formatted (e.g., conversion from binary to ASCII format) and then single nucleotide polymorphisms (SNPs) were called using VarScan (v2.3.7) (*Li et al., 2009*; *Koboldt et al., 2012*). The "pileup2snp" program from VarScan was used with a minimum minor allele frequency threshold of 1%, a read depth of 8, and a minimum of two supporting reads for variant calls. Indels were excluded.

To identify hypervariable loci, we used a geometric distribution based statistic approach as described previously (*Zheng et al., 2010*), which, compared to sliding window searches and other similar methods, has the advantage of avoiding boundary difficulties and variations within contigs. We used a geometric distribution to model the probability of achieving two SNPs separated by a specified non-SNP nucleotide distance. Each between-SNP distance was associated with a probability and the probability of a particular distance occurring by randomly sampling was less than 5%. Thus, we identified a range of SNP distances as significantly less than background if they occurred within our dataset less than 5% of the time.

Protein family domain identification within hypervariable loci ORFs

Protein family domains were identified in ORFs that contained hypervariable loci. The subset of translated virus ORFs that contained hypervariable loci were aligned to the standard Pfam protein family domain database using hmmscan within the HMMer toolkit (version 3.1) and GA gathering bit score thresholds (*Finn, Clements & Eddy, 2011*).

Prediction of single amino acid variant effect on phenotype

The SuSPect algorithm was employed to predict the likelihood of SNP-associated single amino acid variants (SAVs) impacting phenotype. We used SuSPect to create a matrix of likelihood scores for every possible SAV at every position in the ORFs that contained hypervariable loci. This matrix was used as a reference to quantify the likelihood of each hypervariable loci SNP to impact the resulting phenotype. The significance of the score differences between viruses was calculated using a Wilcoxon rank-sum test.

Evolutionary pressure of hypervariable loci and virus genomes

We assessed the evolutionary pressure of a gene using the *pN*/*pS* ratio as in Formula 1, where M_N and M_S represent the observed number of non-synonymous and synonymous SNPs, respectively. These values were normalized by the total number of possible non-synonymous or synonymous substitutions (N_i and S_i, respectively), in order to avoid potential codon usage bias. Furthermore, to normalize for sequence coverage of the SNPs and prevent extreme values, a pseudocount value of an arbitrarily small number was added to the M_N and M_S values, which was defined as half of the square root of the median sequence coverage of SNPs within the region of interest (C_M). The pseudocount approach was used to prevent infinite and illegal values when M_N or M_S had zero values, thus allowing consideration of otherwise infinite or ignored data points. For example, an absence of synonymous mutations would result in an infinitely large value (dividing by zero) thus forcing exclusion of the data point. Our approach preserves this data and allows us to draw conclusions from the largest possible dataset, and has been shown to be effective in previous studies (*Novaes et al., 2008*; *Bajgain et al., 2011*).

$$\frac{pN}{pS} = \frac{\left(\frac{M_N + \frac{\sqrt[2]{C_M}}{2}}{\sum_{i=1}^{L} \frac{N_i}{3}}\right)}{\left(\frac{M_S + \frac{\sqrt[2]{C_M}}{2}}{\sum_{i=1}^{L} \frac{S_i}{3}}\right)} \quad (1)$$

The formula used to calculate the *pN*/*pS* ratio for a gene. M_N is the number of non-synonymous SNPs within the gene and M_S is the number of synonymous SNPs found within the gene. Each mutation value is normalized for the likelihood that the result would have happened by chance, calculated as the sum of the proportions of nucleotides that would have resulted in either a non-synonymous or synonymous mutation. To calculate this proportion, the possible non-synonymous mutations (N) and synonymous mutations (S) at position i within the gene are expressed as a fraction of the three possible alternate nucleotides. SNP counts were smoothed as pseudo-counts using the median SNP sequence coverage (C_M).

This analysis is similar to the *dN*/*dS* calculations often performed to estimate degrees of natural selection among genomes (*Nishida, Frith & Nakai, 2009*; *Schloissnig et al., 2013*).

It is important to note however that such an analytical approach would be inappropriate for this type of sample set because the nucleotide variations are not assignable to isolated strains, which prevents haplotype identification that is a necessary component of *dN/dS* calculations. The *pN/pS* calculation does not assume haplotypes, and is therefore appropriate for metagenomic datasets.

To estimate the selective pressure on hypervariable loci, the locations of the hypervariable loci were extracted, along with their immediately adjacent regions, using a Perl script as presented in the supplemental source code. Adjacent regions are defined as the genomic regions that are two times the length of the hypervariable loci and located immediately before and after the hypervariable loci. These positions were used with the contig sequences, SNP call data, and *pN/pS* calculator to estimate their selective pressure. Hypervariable regions outside of coding regions were not considered.

Calculation of the overall selective pressure on virus contigs was performed in a similar approach to the hypervariable loci selective pressure. Predicted ORFs were first extracted from the contigs using the Glimmer3 toolkit (v3.02) (*Delcher et al., 2007*). The predicted ORFs, along with the contig sequences, SNP profile, and *pN/pS* calculator were used to calculate the overall selective pressure on each gene within each contig. The distributions of selective pressures observed for each gene were observed as categorized by virus type.

Amino acid frequency, charge, and polarity

Amino acid abundance profiles were calculated while correcting for the random probability of that substitution. More specifically, each value was weighted for the number of nucleotides that result in the same amino acid as weighted value = $((number\ of\ nucleotide\ substitutions\ resulting\ in\ same\ amino\ acid)/3)^{-1}$. Relative abundance was calculated as the sum of the corrected frequencies. Charge and polarity were determined using a simple table of known amino acid properties. Differences in profiles between viruses were calculated using a chi-square test.

Diversity generating retroelement identification

We identified potential DGRs by collecting assembled contigs that contained ORFs similar to known reverse transcriptase genes, and a duplicated nucleotide region less than 150 bp in length. Reverse transcriptase ORFs were identified using blastx (e-value $< 10^{-5}$) and the Uniprot reference reverse transcriptase sequences (http://www.uniprot.org/uniprot/?sort=score&desc=&compress=yes&query=%22reverse%20transcriptase%22%20(phage%20OR%20virus)&fil=&format=fasta&force=yes). Repeat regions were identified by comparing each contig to itself with tblastx (e-value $< 10^{-50}$) and were filtered using custom scripts to remove duplicates and regions longer than 150 bp. DGR candidates were removed if they contained no hypervariable loci or if the variable region was not within a predicted ORF.

Diversity generating retroelements were visualized in the Integrated Genomic Viewer using the DGR cassettes and bowtie2 aligned sequences described above. The linkage disequilibrium was calculated using a custom Perl script for formatting and the

"LDheatmap" and "genetics" R packages for analysis and visualization (*Shin et al., 2006*; *Warnes et al., 2003*). The linkage disequilibrium for each pair of SNPs was calculated as the squared allelic correlation (R^2).

Comparison of primary analysis to validation dataset

Near identical contigs were identified between the primary and secondary validation dataset by aligning the two individually assembled contigs to each other using bowtie2, with a specified seed length of 25 and up to one seed mismatch. Sequences from our primary dataset and the Oh et al. *(2014)* dataset were aligned to the near identical contigs. These alignments were used to identify shared SNP locations between our dataset and the Oh et al. *(2014)* dataset. We quantified shared SNP location as percent of our primary analysis SNPs whose location was identical to those of SNPs in the secondary validation dataset. As a control, we compared these results to a simulated dataset where SNP position was randomly assigned. The example SNP alignment over the circular contig was generated using Genious (*Kearse et al., 2012*).

RESULTS

Diversity of skin viruses

We evaluated genomic variability associated with dsDNA skin viruses using a previously published human skin virome metagenomic dataset, consisting of 260,714,906 high quality sequences assembled into >76,000 contigs from 16 individuals (SRA accession: SRP049645) (*Hannigan et al., 2015*). We relied on database virus annotation to identify the taxonomic groups whose contigs had the overall highest confidence matches to reference genomes. Because greater sequencing coverage allows for more refined detection of variable nucleotides (*Schloissnig et al., 2013*), we focused our analysis on taxa whose de novo assembled contigs had sufficient coverage (greater than 10×). Contigs meeting these criteria were identified as *Propionibacterium* phages (contig count = 45), *Staphylococcus* phages (contig count = 319), and human papillomaviruses (HPVs; contig count = 56; Fig. 1A), representing a total of 420 contigs to be used out of the more than 76,000 total contigs in the study (Fig. S1). All of the contigs from these three taxa were used in our analysis, including those below our coverage threshold, since contigs can have regions of high coverage despite an average low coverage. More specific, secondary filtering was done while identifying SNPs. Some contigs were identified as *Pseudomonas* phages or *Enterobacteria* phages, but these taxa excluded from the analysis because their annotations were lower confidence and contig representation was minimal (Fig. 1A). The uneven virus population coverage is potentially a reflection of our inability to taxonomically identify the majority of virome sequences, and as a result we lose this information to the viral "dark matter" (*Hannigan et al., 2015*).

We evaluated the diversity of the skin viruses by constructing phylogenic trees based on conserved viral genes. Only fully assembled ORFs were considered of >1.5 kb for HPV, and >0.15 kb for *Staphylococcus* and *Propionibacterium* phages. Similar to previous studies, we used the L1 major capsid gene to classify HPV strains (*de Villiers et al., 2004*; *Ma et al., 2014*). The terminase large subunit gene was extracted from *Staphylococcus*

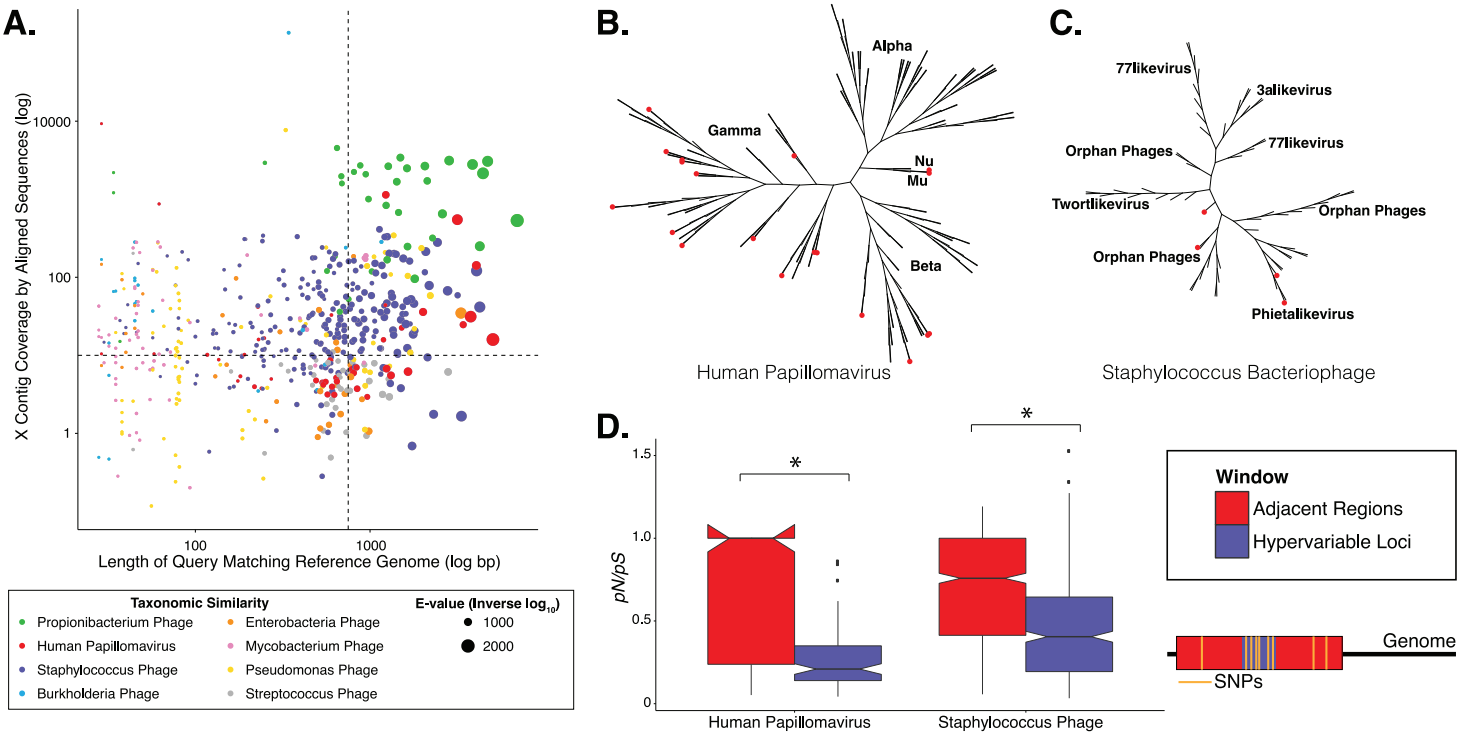

Figure 1 Phylogenetic & evolutionary characteristics of skin virome hypervariable loci. (A) Scatter plot depicting the candidate contigs considered for analysis in this study. Each point is a contig that mapped to a reference virus genome. The *x*-axis shows the length ($\log_{10}$ scale) of the contig subsection that mapped to the reference genome. The *y*-axis shows the overall coverage of the contig, as a quantification of sequences aligning to the contig. The color highlights the reference virus genome that the contig was most similar to, and the size depicts the *e*-value (inverse $\log_{10}$) associated with the contig-reference match. The horizontal dashed line marks the threshold of 10× coverage, and the vertical dashed line marks the 750 bp length threshold. (B) Phylogenetic tree of skin virome HPVs and (C) *Staphylococcus* phages, structured onto a standard phylogenetic tree using reference genomes. HPV phylogeny was based on the L1 major capsid gene and *Staphylococcus* phage phylogeny was based on the large terminase subunit. Contigs from this study are highlighted as orange dots, and genera are labeled with text. Phylogenetic lengths were normalized to ranks to facilitate visualization. (D) Box plots depicting the evolutionary pressure of HPVs (left) and *Staphylococcus* bacteriophages (right) at the hypervariable loci (blue) and the regions immediately adjacent to the hypervariable loci (red). Adjacent regions were calculated as being twice the length of the hypervariable loci (see visualization to the right). The hypervariable locus and adjacent region (combination of both sides) from each sample were evaluated for evolutionary pressure (*y*-axis) using SNPs (pink lines in right illustration). Asterisk indicates a statistically significant difference ($p < 0.01$). Notched boxplots were created using ggplot and show the median (center line), the inter-quartile range (IQR; upper and lower boxes), the highest and lowest value within 1.5 × IQR (whiskers), and the notch which provides an approximate 95% confidence interval as defined by 1.58 × IQR/sqrt(*n*).

phage contigs to construct phylogeny as described previously (*Gutiérrez et al., 2013*; *Ma et al., 2014*). Because this gene is used for phylogeny of a variety of phages, we attempted to construct *Propionibacterium* phage phylogeny in a similar manner (*Ganz et al., 2014*; *Li et al., 2014*), but were ultimately unsuccessful due to the lack of a full-length de novo assembled reference genes in the dataset.

Most skin HPVs were identified as gamma HPVs, the prototypical cutaneous HPV class (Fig. 1B) (*Mistry, Wibom & Evander, 2008*). Few contigs were identified as beta and Mu/Nu HPVs, and none were identified as alpha HPVs. This is consistent with data from the Human Microbiome Project cohort (*Ma et al., 2014*).

Fewer *Staphylococcus* phage marker genes were identified, compared to HPVs, likely because *Staphylococcus* phage genomes are orders of magnitude longer than HPV

genomes, thereby decreasing the probability that contigs covered the entire genome. Because multiple displacement amplification (MDA) was not used to create this dataset, there is no MDA-associated bias toward small circular genomes. The *Staphylococcus* phage contigs belonged to the Phietalikevirus genus and orphan virus groups (those that have not yet been classified) (Fig. 1C). Of the *Staphylococcus* phage contigs identified, 49.6% (123 out of 248 contigs) were predicted to be lysogenic, based on similarity to lysogenic phages in the ACLAME database, integrase genes in the Uniprot database, and Staphylococcus reference genomes from the NCBI nucleotide database, as described previously (*Leplae, Lima-Mendez & Toussaint, 2010*; *Minot et al., 2011*; *Hannigan et al., 2015*). This is a minimum estimate of contig lysogeny, as some of the other contigs may have lysogenic signatures that we failed to identify. Furthermore, because this classification strategy is based on blast assignments, it may result in false positives if genes in the database are homologous to genes present in lytic phages.

Hypervariable loci within the skin virome

We implemented a geometric distribution-based approach to identify regions of high genomic diversity, as in (*Zheng et al., 2010*). Regions within each contig that contained a significantly higher frequency of SNPs over the stochastic background were identified as viral hypervariable loci. Significance was defined as the frequency of SNPs having less than a 5% chance of randomly occurring, given the geometric distribution of the dataset. HPVs and *Staphylococcus* phages maintained 106 and 465 hypervariable loci, respectively. We were unable to detect hypervariable loci in the *Propionibacterium* phage population.

To determine the virus protein family domains hosting hypervariable loci, we used the hidden Markov model analysis implemented by HMMer (*Finn, Clements & Eddy, 2011*). Hypervariable loci-containing HPV genes include E6, E2, and E1 genes, which are associated with infectious gene expression, and the L1 major capsid gene, which is involved in tropism and host immune evasion (Table S1). The L1 major capsid protein is also a target in contemporary, widely used HPV vaccines (*Schiller & Lowy, 2012*). Hypervariable loci were detected in a variety of *Staphylococcus* phage genes with predicted functions related to tropism, host immune evasion, and utilization of host resources (Table S2).

Selective pressures on hypervariable loci

We evaluated the selective pressures on virus genes by calculating the *pN*/*pS* ratio of non-synonymous SNPs (*pN*) to synonymous SNPs (*pS*) within each virus taxa (*Schloissnig et al., 2013*). This was used as an alternative to *dN/dS* because *dN/dS* assumes haplotype information which cannot be fulfilled by metagenomic data (*Schloissnig et al., 2013*). In the *pN*/*pS* calculation, neutral evolution is defined as an equal frequency of synonymous and non-synonymous polymorphisms. Selective pressure favors non-synonymous mutations, resulting in increased *pN*/*pS* ratios. Purifying selection has the opposite effect. Because the existing model (*Schloissnig et al., 2013*) is susceptible to stochastic effects and extreme outliers (e.g., division by zero when $pS = 0$), we added a pseudocount correction (Formula 1).

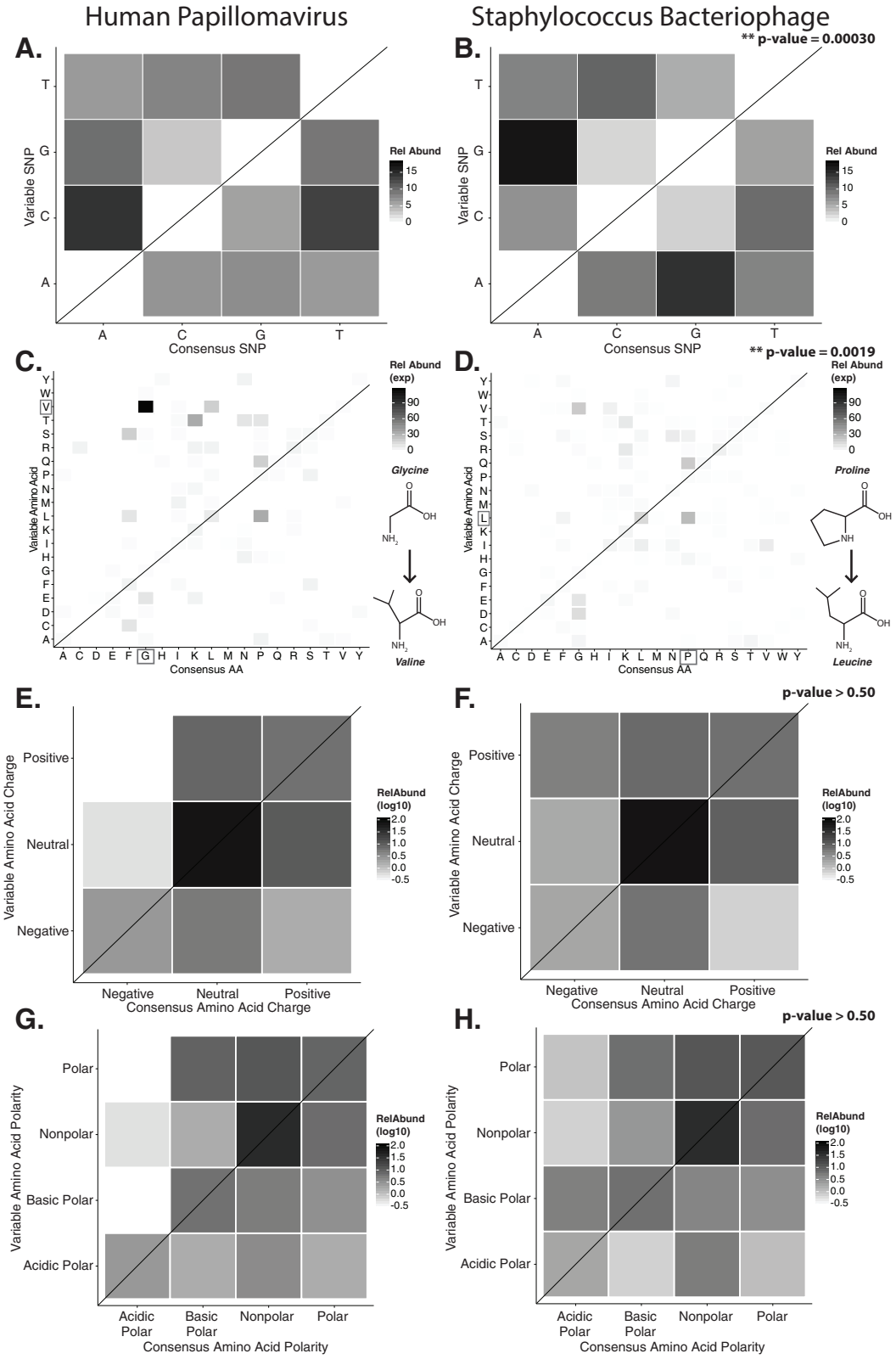
Human Papillomavirus
Staphylococcus Bacteriophage
A.
B.
** p-value = 0.00030
Variable SNP
Consensus SNP
Rel Abund
C.
D.
** p-value = 0.0019
Variable Amino Acid
Consensus AA
Rel Abund (exp)
Glycine
Valine
Proline
Leucine
E.
F.
p-value > 0.50
Variable Amino Acid Charge
Consensus Amino Acid Charge
RelAbund (log10)
Positive
Neutral
Negative
G.
H.
Variable Amino Acid Polarity
Consensus Amino Acid Polarity
Polar
Nonpolar
Basic Polar
Acidic Polar

Figure 2 Nucleotide and amino acid substitution patterns within viral hypervariable loci. Heat maps portraying the counts of every possible nucleotide substitution for each SNP found within (A) HPV and (B) *Staphylococcus* phage hypervariable loci. Tile color weight corresponds to the relative abundance of SNP substitution counts. The diagonal line highlights the panels associated with no substitution. The substitution patterns of amino acids at each SNP are also shown with exponential transformation (C, D). An illustration of the major amino acid substitutions are provided beneath the legends as a reference. Amino acid charge (E, F) and polarity with acidity (G, H) are shown with $\log_{10}$ transformation. The absence of a basic or acidic polar identifier indicates the amino acid 20 is polar but neutral. The HPV substitution profiles are found in the left column and the *Staphylococcus* phage profiles are found on the right. Chi-square significance *p*-value, comparing variation profiles between the viruses in each row (i.e., A and B), is shown in the upper right corner of the associated *Staphylococcus* phage variation profile. The most frequently substituted amino acid pairs are highlighted with a box around the amino acid letters.

We determined whether hypervariable loci are in fact loci of focused selective pressure by comparing *pN*/*pS* values of the loci to the adjacent genomic regions. *pN*/*pS* values of hypervariable loci were significantly lower than adjacent regions in both HPV (median: adjacent = 1.0, hypervariable loci = 0.21; *p*-value = 3.4e−17) and *Staphylococcus* phage (median: adjacent = 0.76, hypervariable loci = 0.41; *p*-value = 1.8e−40) genomes, suggesting purifying selection and a propensity to maintain existing protein sequences (Figs. 1D). HPV hypervariable loci were under significantly greater purifying selection than those of *Staphylococcus* phages (median: HPV = 0.21, *Staphylococcus* phage = 0.41; *p*-value = 4.64e−9) (Fig. S1). Furthermore, not only are the *pN/pS* values of the hypervariable significantly lower than their adjacent regions, but very few of the loci have a *pN/pS* value greater than one.

To evaluate whether the observed selective pressure in HPV and *Staphylococcus* virus communities is genome-wide or localized to hypervariable loci, we quantified the selective pressure on each virus' genome by calculating the overall *pN*/*pS* ratio including hypervariable loci and non-hypervariable loci SNPs. We observed nearly neutral pressure across HPVs and *Staphylococcus* phages that mirrored pressures to those observed in the regions adjacent to the hypervariable loci (median: HPV = 1.0, *Staphylococcus* phage = 0.81, *p*-value = 3.2e−5) (Fig. S2).

Functional implications of targeted substitutions within hypervariable loci

In order to evaluate the specific nucleotide changes occurring at hypervariable loci, as well as to evaluate the implications of specific nucleotide polymorphisms, we quantified the frequency of individual nucleotide substitutions within hypervariable loci. A>C and T>C substitutions were most frequent in HPV hypervariable loci (Fig. 2A). *Staphylococcus* phages exhibited a significantly different substitution profile (*p*-value = 0.00018, chi-square test), with the most common substitutions being A>G and G>A transitions (Fig. 2B). HPV and *Staphylococcus* phage substitutions were more likely to be transitions, with a transition/transversion (ti/tv) ratio of 3.25 and 2.02, respectively.

We predicted how hypervariable loci SNPs might affect protein functionality by evaluating patterns of the amino acid substitutions while correcting for the random chance that the substitution will occur. The most frequent non-synonymous amino acid

substitution in HPVs was glycine (consensus amino acid) to valine (variant amino acid, Fig. 2C). While these amino acids are (non-polar and hydrophobic), glycine is less hydrophobic than valine. The most frequent non-synonymous amino acid substitution in *Staphylococcus* phages was proline to leucine (Fig. 2D), a substitution between a non-polar cyclic amino acid and an aliphatic straight chain amino acid. Profiles of amino acid substitution were significantly different between HPVs and *Staphylococcus* phages (p-value = 0.0021; chi-square test).

Amino acid polarity and charge were largely maintained in HPV hypervariable loci (Figs. 2E and 2G). In instances of altered charge, visual inspection suggests the most frequent changes were from neutral to positive or negative charge, or positive to neutral charge. Consensus acidic polar residues were not associated with polymorphisms. *Staphylococcus* phage community hypervariable loci appeared to be under weaker substitution selection, with a greater diversity in amino acid charge and polarity (Figs. 2F and 2H) compared to HPV. Patterns of substitution charge and polarity were not significant (p-value > 0.5; chi-square test) when comparing the entire HPV to *Staphylococcus* phage substitution profiles.

We reinforced the observed functional implications of hypervariable loci by predicting the effects of their associated SAVs on gene phenotype using the support vector machine algorithm implemented in SuSPect (*Yates et al., 2014*). This method assigns a deleterious score to each hypervariable loci SNP-associated SAV, with 0 representing a neutral SAV and 100 representing a SAV with high likelihood to impact phenotype. These scores are based on the predicted impact of the SAV on the tertiary and secondary structure of the resulting protein, the location of the SAV within the resulting protein (surface vs core), and whether the SAV has previously been associated with altered protein–protein interactions. Both *Staphylococcus* phages and HPVs have an abundance of SNPs associated with SAVs predicted to be deleterious (deleterious scores approaching 100) (Fig. 3). The HPV SNPs were predicted to be significantly more likely to impact phenotype than the *Staphylococcus* phage SNPs (median: HPV = 45, *Staphylococcus* phage = 17; p-value < 2.2e−16), suggesting that SNPs impact functionality differently between viruses.

Diversity generating retroelements as a mechanism for targeted hypervariability

Diversity generating retroelements are a genetic system used by bacteriophages (as well as bacteria and archea) to promote targeted hypervariability in genes (*Doulatov et al., 2004*). While DGRs are complex and consist of many components, at their most basic they can be identified as elements consisting of a reverse transcriptase gene and a repeated nucleotide sequence of length <150 bp that is found in two separate locations of the genome (*Doulatov et al., 2004*; *Minot et al., 2012*; *Schillinger et al., 2012*), termed the template region and the variable region. The template region is transcribed, then reverse transcribed in an error-prone fashion. The resulting cDNA is then integrated into the variable region, introducing base substitutions. Targeted hypervariation impacts functions including broadened host cell tropism by mutagenizing a phage tail fiber gene (*Doulatov et al., 2004*).

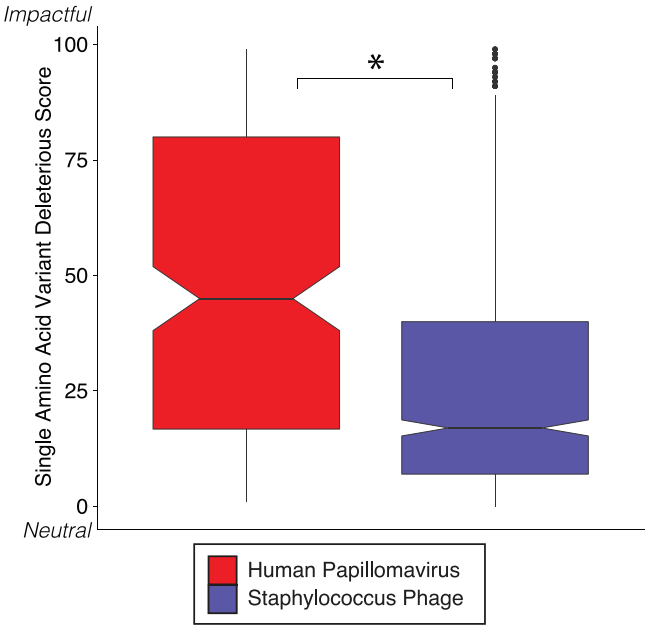

Figure 3 SVM predicted impact of hypervariable loci on phenotype. Notched boxplot of deleterious scores in human papillomavirus (red) and *Staphylococcus* phage (blue) genomes. A low deleterious score indicates a predicted neutral phenotypic effect, while a high score indicates a predicted strong phenotypic effect. Asterisk indicates significant difference by Wilcoxon rank-sum test ($p < 1e15$). Boxplot parameters as described in Fig. 1.

We thus sought to identify candidate DGR cassettes within our viral contigs. We defined the candidate cassettes as pairs of non-overlapping regions with similar nucleotide sequences (tblastx of contigs against themselves, *e*-value < 10e–50) and co-localized on a contig containing a predicted virus/phage reverse transcriptase gene. We only considered cassettes that were located within a predicted viral gene, contained at least one hypervariable locus in their variable region, and exhibited truly random variation (different between reads). Based on these criteria, we identified one *Staphylococcus* phage DGR candidate that contained hypervariable loci. We also identified five other DGR candidates that were associated with hypervariable loci outside of predicted genes or that failed to demonstrate linkage disequilibrium, suggesting an association with cryptic genes or pseudogenes.

For the *Staphylococcus* phage DGR candidate with hypervariable loci, we calculated the linkage disequilibrium associated with the variable nucleotide positions to infer whether the DGR was active or inactive (e.g., an evolutionary artifact). The DGR cassette had unlinked nucleotide variation, which was supported by low levels of linkage disequilibrium (squared allelic correlation R^2) between SNP pairs in the variable region (Fig. 4). In this cassette, the template region has less frequent blocks of linkage equilibrium (unlinked variants) while the variable region was associated with greater linkage equilibrium. Together this suggests the observed *Staphylococcus* phage is active. The variable region was associated with a gene of unknown function.

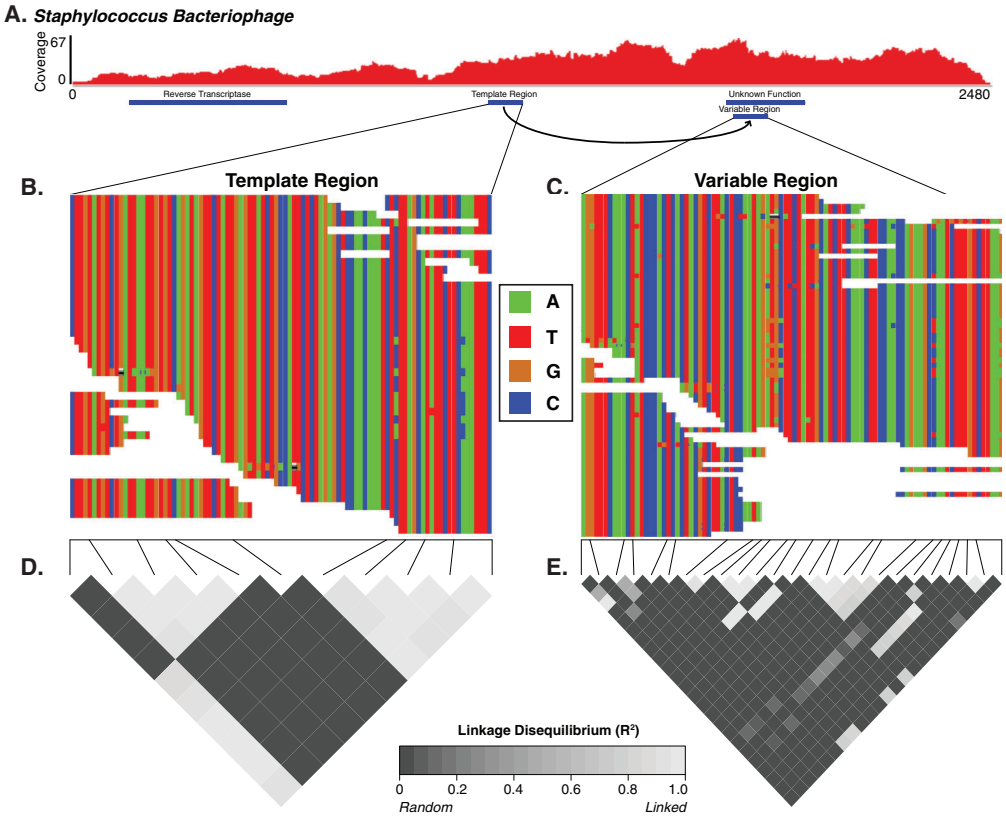

Figure 4 The diversity generating retroelement as a mechanism for targeted nucleotide variation. Alignment illustrating a putative diversity generating retroelement in *Staphylococcus* phage. (A) Sashimi plot of sequence coverage across the contig. Coverage ranges from 0 to 67×. Below the coverage is a map of the relevant genes predicted within the contig. Sequence alignment of the diversity generating retroelement template region (B) and variable region (C). Linkage disequilibrium heatmap for the template (D) and variable (E) region. Panels compare variable nucleotides to each other and darker tiles indicate decreased linkage disequilibrium correlation, according to squared allelic correlation (R^2) between pairs of SNPs.

Skin virome variability patterns and SNP locations are reproducible across different datasets

We repeated our analyses in a separate, independently collected dataset from another research group (SRA BioProject: 46333) (*Oh et al., 2014*) to determine the generalizability of our findings. We analyzed metagenomic sequence data of skin specimens that were collected from the retroauricular crease without initial purification of virus like particles. Consistent with our primary analysis, *Staphylococcus* and *Propionibacterium* phages were identified as having the highest coverage and similarity to reference genomes (Fig. 5A). *Pseudomonas* phages were identified but were in the minority and had low coverage and similarity to reference genomes. HPV was not identified as a major virus in our analysis of the retroauricular crease; however, molluscum contagiosum virus, a poxvirus that causes cutaneous growths that become severe in immunocompromised states, was present in high relative abundance in agreement with the original published findings (*Oh et al., 2014*).

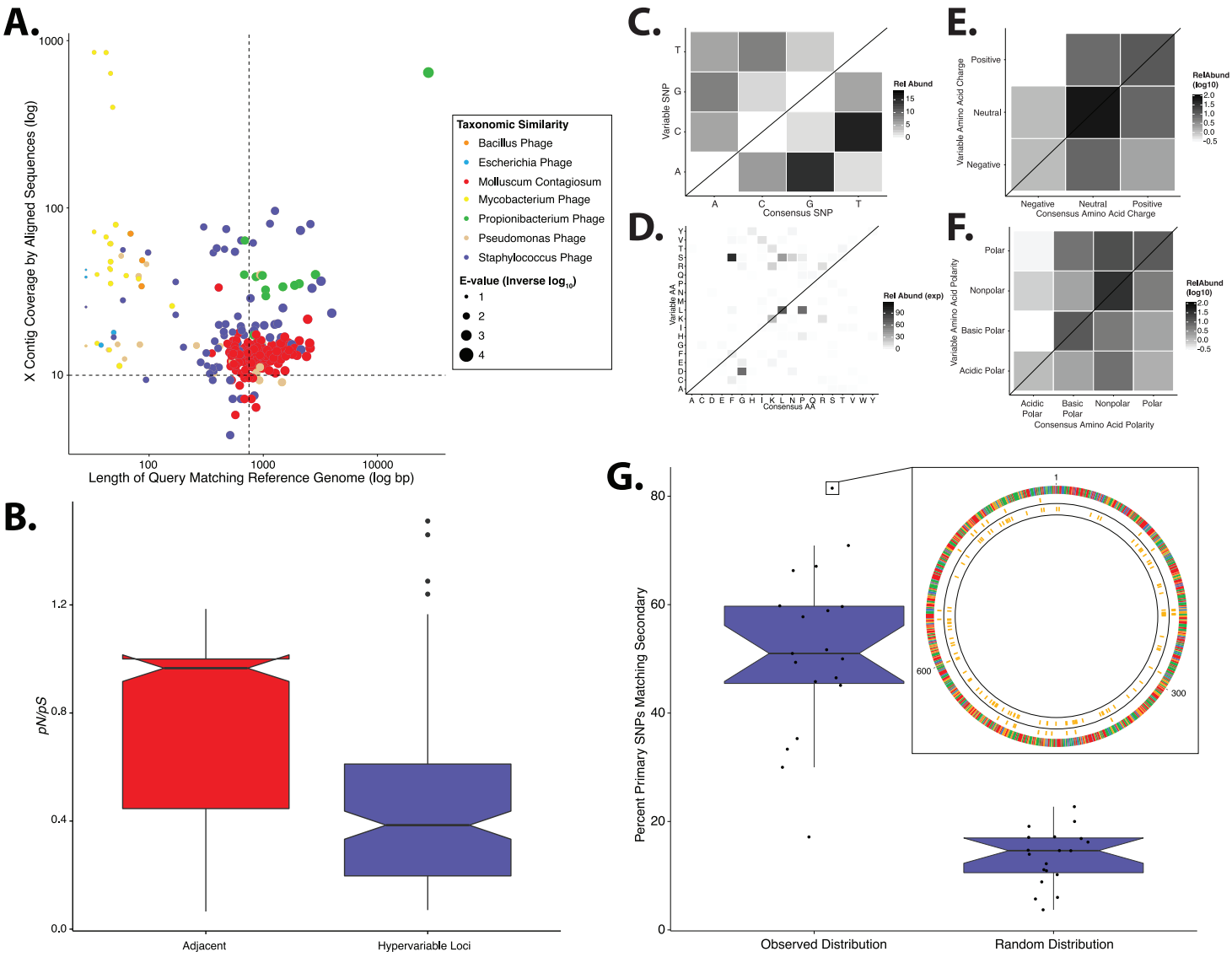

Figure 5 Validation of study findings using secondary dataset. Results from the Oh et al. *(2014)* dataset, which was analyzed using the same workflow as the primary dataset. (A) Scatter plot depicting the candidate contigs considered for analysis in this study. Each point is a contig that mapped to a reference virus genome. The *x*-axis shows the length (in nucleotides) of the contig subsection that mapped to the reference genome. The *y*-axis shows the overall coverage of the contig as a quantification of sequences aligning to the contig. The color highlights the reference virus genome that the contig was most similar to, and the size depicts the blast bit score associated with the contig-reference match. (B) Box plots depicting the evolutionary pressure of *Staphylococcus* bacteriophages at the hypervariable loci (blue) and the regions immediately adjacent to the hypervariable loci (red). (C) Heat map portraying the counts of every possible nucleotide substitution for each SNP found within 21 *Staphylococcus* phage hypervariable loci. Tile color weight corresponds to the relative abundance of SNP substitution counts. The diagonal line highlights the panels associated with no substitution. The substitution patterns of amino acids at each (D) SNP, (E) amino acid charge, and (F) polarity with acidity are also shown. (G) Notched boxplot illustrating the percent of primary dataset SNPs whose nucleotide positions were identical to those from the secondary validation sample set (left) compared to a simulated dataset of randomly assigned SNP locations (right). The inset shows an example contig identified in both datasets with 81% identical SNP positions. SNPs are represented as yellow lines, with the inner circle representing the validation dataset, and the middle circle representing the primary dataset. The outmost ring illustrates the contig, colored by nucleotides (A = red, C = blue, G = yellow, T = green). Boxplot parameters as described in Fig. 1.

Similar to our primary analysis, we identified 158 hypervariable loci within the *Staphylococcus* phage communities, and observed only 12 hypervariable loci associated with *Propionibacterium* phages, further highlighting the overall lack of genetic variability

of the *Propionibacterium* phage communities. The *Staphylococcus* phage hypervariable loci were associated with purifying selection, yielding a *pN/pS* ratio slightly below 0.4 (Fig. 5B), recapitulating our findings in the primary dataset.

Staphylococcus phage nucleotide substitutions were associated with transitions between guanine and adenine residues, as we observed in our primary analysis (Fig. 5C). The ti/tv ratio of these loci was 1.02. The most common amino acid substitution was proline to leucine (Fig. 5D) and the substitution properties appeared loosely specific based on charge and polarity (Figs. 5E and 5F), reproducing our findings (Fig. 2).

We also evaluated the reproducibility of SNP position between identical but independently assembled genomic contigs of the two studies. We quantified the proportion of SNPs in our primary dataset that were also found at the same position in the secondary dataset. This revealed a median of approximately 50% overlap between datasets (Fig. 5G). As a control, we generated a simulated dataset using randomly assigned SNP positions instead of those determined experimentally. This yielded a significantly lower median of approximately 15% shared nucleotide SNP calls (Fig. 5G), suggesting that the observed SNP position is not random. These data indicate that our findings are consistent across different skin virome populations and techniques of collection and sequencing.

DISCUSSION

Here we report localized targeted hypervariability in some of the most prevalent members of the skin virome. Hypervariable loci provide a substrate for complex virus evolution throughout the virome, which manifest as natural selection that differs by virus type and enforces purifying selection. Hypervariable loci, which were present in genes encoding factors including virus tropism and host immune evasion, and were primarily under purifying selection, whereas overall virus genomes were under near neutral selection. We characterized selected substitution of nucleotides within hypervariable loci, with different variant patterns between HPVs and *Staphylococcus* bacteriophage communities. These findings were validated in an independently collected cohort.

We showed *Propionibacterium* phages exhibited strikingly low nucleotide variation with nearly no identifiable hypervariable loci. While this starkly contrasts with HPVs and *Staphylococcus* bacteriophages, it agrees with our current understanding of *Propionibacterium* phage diversity. Genome comparisons of *Propionibacterium* phage isolates revealed minimal nucleotide diversity, although this has yet to be supported by targeted metagenomic evidence such as presented here (*Marinelli et al., 2012*; *Liu et al., 2015*). The lack of *Propionibacterium* phage hypervariability in our metagenomic dataset provides another level of evidence for minimal *Propionibacterium* phage diversity on the skin.

There are several potential factors that could contribute to the limited diversity of *Propionibacterium* phages, and a consensus has yet to be reached. The lack of hypervariable loci suggests minimal evolutionary pressure on the phages, which may be a reflection of their environment. As suggested previously, the phages and their hosts reside in a unique and relatively isolated environment deeper in the skin, which may contribute to

low genomic diversity (*Marinelli et al., 2012*). Our data further support this hypothesis. Another factor that could contribute to differential phage genomic diversity is their host range. Although *Propionibacterium* phages have broad infectious capabilities within bacterial species, they may be limited in their ability to infect other species (*Marinelli et al., 2012*). *Staphylococcus* phages demonstrate greater genomic diversity, and may be capable of infecting a broader range of hosts.

We observed greater selective pressure on HPVs compared to *Staphylococcus* phages, which may reflect greater pressures from the human immune system, compared to phage bacterial hosts. This may also reflect the effects of different virus replication cycles on evolutionary selection. HPVs do not usually exist in a latent, integrated state, while *Staphylococcus* phages do (*Bae et al., 2006*; *Goerke et al., 2009*; *Edwards et al., 2013*). Our data suggest that at least one-half of the observed *Staphylococcus* phages have temperate replication cycles. As long as the *Staphylococcus* phages are integrated into the bacterial genome, we hypothesize that they are under less selective pressure by external factors.

The viral hypervariable loci are primarily associated with purifying selective pressure, a finding in agreement with previous non-metagenomic virus reports (*Chen et al., 2005*; *Wolf et al., 2006*; *Li et al., 2011*). The observed prominent purifying selective pressure supports an evolutionary model of long static periods punctuated by brief positive selection, as is observed in influenza virus (*Wolf et al., 2006*). Here nucleotide diversity acts as a primer for rapid virus adaptation through brief positive selection, while maintaining periods of consistency through purifying selection during static environmental conditions. As an example, some localized nucleotide diversity may allow for the generation of phages with different tropisms (e.g., different bacterial strains). If there are limited hosts, the phages that successfully infect those hosts will be selected for, and altered tropisms will be actively selected against. If that host population changes, then those viruses with the appropriate tropism will be selected for instead of being selected against. Ultimately, longitudinal and strain specific studies will be required to further address this hypothesis.

The amino acid substitutions associated with hypervariable loci were non-random and followed virus-specific substitution patterns (Fig. 2). HPV hypervariable loci were most associated with substitutions from glycine to valine. This substitution has recently been associated with infectious functionality, whereby the introduction of this mutation resulted in impaired infective ability of the virus (*Bronnimann et al., 2013*). This impaired infectious activity was attributed to a reduced efficacy of genomic DNA endosomal translocation within the host, which may have been the result of impaired trans-membrane alpha-helical self-association of the L2 minor capsid protein. Given these findings, our results suggest hypervariable loci are involved in promoting diversity in endosomal translocation motifs to some degree. Hypervariable loci may certainly have other diverse, functional roles, as evidenced by the wide range of hypervariable loci-containing genes.

The dominant amino acid substitution observed in *Staphylococcus* bacteriophages was from proline to leucine, a different substitution than that observed in HPV. This substitution could affect protein structure, particularly a loss of rigidity due to the loss of the proline ring structure. This observation may reflect a biologically important

adaptation of the bacteriophage to its *Staphylococcus* host, which have been shown to be auxotrophic for proline and leucine and may switch between auxotroph and prototroph depending on nutrient availability (*Emmett & Kloos, 1975*; *Nuxoll et al., 2012*). Because the amino acids may be in variable supply depending on the host, phages may alter their amino acid usage to exploit what is most readily available.

The overall selective nucleotide substitutions associated with HPV amino acid charge highlights a potential maintenance of HPV tropism. The lack of HPV substitutions between charges may suggest a selection against strong alterations in protein isoelectric points, which have been implicated in affecting HPV tropism (*Mistry, Wibom & Evander, 2008*). Furthermore, because acidic residues almost never mutated to non-polar residues, these acidic amino acids are potentially important external amino acids that may participate in tropic protein–protein interactions.

The described patterns in our findings suggest a role for targeted and/or localized genomic variation. One mechanism of such active targeted variation in *Staphylococcus* bacteriophages is DGRs. In this system, a phage-encoded reverse transcriptase copies a template region to create a variable region in a gene in an error-prone fashion. We identified such an element that is likely active and promotes diversity in a gene of unknown function. We additionally identified five other DGR elements whose hypervariable loci were not associated with an identified gene, suggesting an interesting phenomenon where high variability is selected for in non-coding regions. While informative, these discoveries only explain the diversity-generating mechanism of a small proportion of hypervariable loci. We suspect another underlying mechanism for the origin and evolution of other hypervariable loci could be that they are located on functionally important loci such as encoding regions that interact with other genes or are important to protein structure, therefore being functionally selected. Significant further investigation will be needed to characterize these and other potential mechanisms behind the observed hypervariable loci.

This study illustrates the diversity of evolutionary pressures on skin virus communities. It begins to provide further community-wide context to the molecular understanding of skin viruses, and highlights important aspects of their infectious cycles. These insights also contribute to understanding virus ecology of the human skin, and will inform future translational research into HPV vaccination, vaccination against other skin-associated viruses, effects of phages on bacterial pathogenesis, and phage therapy. Understanding how viruses evolve in their natural communities is crucial for improving these translational applications, and our findings here, which focus on HPV and *Staphylococcus* phages, will benefit cutaneous clinical virology and provide a foundation for future studies.

CONCLUSION

We report that the skin virus communities contain hypervariable loci that are associated with strong purifying selection and targeted nucleotide substitution. The degree of selective pressure and impact of amino acid substitutions on protein chemistry (structure, isoelectric point, polarity) is virus specific, despite being members of the same community. These hypervariable loci are found within diverse viral strains, with varying

degrees of phylogenetic divergence over their evolutionary history. We further reproduce these findings in independently collected skin virus communities.

ACKNOWLEDGEMENTS

We thank the members of the Grice and Bushman laboratories for their underlying contributions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by grants from the NIH (NIAMS R00AR060873 to Elizabeth A. Grice and NIAMS R01AR066663 to Elizabeth A. Grice). Geoffrey D. Hannigan is supported by the Department of Defense, National Defense Science and Engineering Graduate fellowship program and Jacquelyn S. Meisel is supported by NIH T32 HG000046 Computational Genomics Training Grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NIH: NIAMS R00AR060873 and NIAMS R01AR066663.
NIH Computational Genomics Training Grant: T32 HG000046.

Competing Interests

Samuel S. Minot is an employee of One Codex.

Author Contributions

- Geoffrey D. Hannigan conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Qi Zheng analyzed the data, reviewed drafts of the paper.
- Jacquelyn S. Meisel analyzed the data, reviewed drafts of the paper.
- Samuel S. Minot analyzed the data, reviewed drafts of the paper.
- Frederick D. Bushman analyzed the data, reviewed drafts of the paper.
- Elizabeth A. Grice conceived and designed the experiments, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Deposition

The following information was supplied regarding data availability:
GitHub, ViromeVarScripts, https://github.com/Microbiology/ViromeVarScripts.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2959#supplemental-information.

REFERENCES

Bacher JM, Bull JJ, Ellington AD. 2003. Evolution of phage with chemically ambiguous proteomes. *BMC Evolutionary Biology* **3**:24 DOI 10.1186/1471-2148-3-24.

Bae T, Baba T, Hiramatsu K, Schneewind O. 2006. Prophages of *Staphylococcus aureus* Newman and their contribution to virulence. *Molecular Microbiology* **62(4):**1035–1047 DOI 10.1111/j.1365-2958.2006.05441.x.

Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA. 2011. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). *BMC Genomics* **12(1):**370 DOI 10.1186/1471-2164-12-370.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology* **13**:R122 DOI 10.1186/gb-2012-13-12-r122.

Borghans JAM, Beltman JB, De Boer RJ. 2004. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* **55(11):**732–739 DOI 10.1007/s00251-003-0630-5.

Bronnimann MP, Chapman JA, Park CK, Campos SK. 2013. A transmembrane domain and GxxxG motifs within L2 are essential for papillomavirus infection. *Journal of Virology* **87(1):**464–473 DOI 10.1128/JVI.01539-12.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10(1):**421 DOI 10.1186/1471-2105-10-421.

Chen Z, Terai M, Fu L, Herrero R, DeSalle R, Burk RD. 2005. Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. *Journal of Virology* **79(11):**7014–7023 DOI 10.1128/jvi.79.11.7014-7023.2005.

Das SR, Hensley SE, Ince WL, Brooke CB, Subba A, Delboy MG, Russ G, Gibbs JS, Bennink JR, Yewdell JW. 2013. Defining influenza A virus hemagglutinin antigenic drift by sequential monoclonal antibody selection. *Cell Host and Microbe* **13(3):**314–323 DOI 10.1016/j.chom.2013.02.008.

de Villiers E-M, Fauquet C, Broker TR, Bernard H-U, zur Hausen H. 2004. Classification of papillomaviruses. *Virology* **324(1):**17–27 DOI 10.1016/j.virol.2004.03.033.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23(6):**673–679 DOI 10.1093/bioinformatics/btm009.

Donaldson EF, Lindesmith LC, Lobue AD, Baric RS. 2010. Viral shape-shifting: norovirus evasion of the human immune system. *Nature Reviews Microbiology* **8(3):**231–241 DOI 10.1038/nrmicro2296.

Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF. 2004. Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431(7007):**476–481 DOI 10.1038/nature02833.

Edwards TG, Helmus MJ, Koeller K, Bashkin JK, Fisher C. 2013. Human papillomavirus episome stability is reduced by aphidicolin and controlled by DNA damage response pathways. *Journal of Virology* **87(7):**3979–3989 DOI 10.1128/JVI.03473-12.

Emmett M, Kloos WE. 1975. Amino acid requirements of staphylococci isolated from human skin. *Canadian Journal of Microbiology* **21(5):**729–733 DOI 10.1139/m75-107.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**:W29–W37 DOI 10.1093/nar/gkr367.

Ganz HH, Law C, Schmuki M, Eichenseher F, Calendar R, Loessner MJ, Getz WM, Korlach J, Beyer W, Klumpp J. 2014. Novel giant siphovirus from *Bacillus anthracis* features unusual genome characteristics. *PLoS ONE* **9(1):**e85972 DOI 10.1371/journal.pone.0085972.

Goerke C, Pantucek R, Holtfreter S, Schulte B, Zink M, Grumann D, Broker BM, Doskar J, Wolz C. 2009. Diversity of prophages in dominant *Staphylococcus aureus* clonal lineages. *Journal of Bacteriology* **191(11):**3462–3468 DOI 10.1128/JB.01804-08.

Guan M, Wang W, Liu X, Tong Y, Liu Y, Ren H, Zhu S, Dubuisson J, Baumert TF, Zhu Y, Peng H, Aurelian L, Zhao P, Qi Z. 2012. Three different functional microdomains in the hepatitis C virus hypervariable region 1 (HVR1) mediate entry and immune evasion. *Journal of Biological Chemistry* **287(42):**35631–35645 DOI 10.1074/jbc.M112.382341.

Guo H, Arambula D, Ghosh P, Miller JF. 2014. Diversity-generating retroelements in phage and bacterial genomes. *Microbiology Spectrum* **2(6):**1237–1252 DOI 10.1128/microbiolspec.mdna3-0029-2014.

Gutiérrez D, Adriaenssens EM, Martínez B, Rodríguez A, Lavigne R, Kropinski AM, García P. 2013. Three proposed new bacteriophage genera of staphylococcal phages: "3alikevirus," "77likevirus" and "Phietalikevirus". *Archives of Virology* **159(2):**389–398 DOI 10.1007/s00705-013-1833-1.

Hannigan GD, Grice EA. 2013. Microbial ecology of the skin in the era of metagenomics and molecular microbiology. *Cold Spring Harbor Perspectives in Medicine* **3(12):**a015362 DOI 10.1101/cshperspect.a015362.

Hannigan GD, Hodkinson BP, McGinnis K, Tyldsley AS, Anari JB, Horan AD, Grice EA, Mehta S. 2014. Culture-independent pilot study of microbiota colonizing open fractures and association with severity, mechanism, location, and complication from presentation to early outpatient follow-up. *Journal of Orthopaedic Research* **32(4):**597–605 DOI 10.1002/jor.22578.

Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ, Minot S, Bushman FD, Grice EA. 2015. The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *mBio* **6(5):**e01578-15 DOI 10.1128/mBio.01578-15.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486(7402):**207–214 DOI 10.1038/nature11234.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4):**772–780 DOI 10.1093/molbev/mst010.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28(12):**1647–1649 DOI 10.1093/bioinformatics/bts199.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22(3):**568–576 DOI 10.1101/gr.129684.111.

Kubinak JL, Ruff JS, Hyzer CW, Slev PR, Potts WK. 2012. Experimental viral evolution to specific host MHC genotypes reveals fitness and virulence trade-offs in alternative MHC types. *Proceedings of the National Academy of Sciences of the United States of America* **109(9):**3422–3427 DOI 10.1073/pnas.1112633109.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Thurber RLV, Knight R, Beiko RG, Huttenhower C. 2013. Predictive functional

profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31(9):**814–821 DOI 10.1038/nbt.2676.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4):**357–359 DOI 10.1038/nmeth.1923.

Leplae R, Lima-Mendez G, Toussaint A. 2010. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research* **38**:D57–D61 DOI 10.1093/nar/gkp938.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16):**2078–2079 DOI 10.1093/bioinformatics/btp352.

Li S, Fan H, An X, Fan H, Jiang H, Chen Y, Tong Y. 2014. Scrutinizing virus genome termini by high-throughput sequencing. *PLoS ONE* **9(1):**e85806 DOI 10.1371/journal.pone.0085806.

Li W, Shi W, Qiao H, Ho SYW, Luo A, Zhang Y, Zhu C. 2011. Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses. *Virology Journal* **8(1):**183 DOI 10.1186/1743-422x-8-183.

Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, Warner BB, Tarr PI, Wang D, Holtz LR. 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature Medicine* **21(10):**1228–1234 DOI 10.1038/nm.3950.

Liu J, Yan R, Zhong Q, Ngo S, Bangayan NJ, Nguyen L, Lui T, Liu M, Erfe MC, Craft N, Tomida S, Li H. 2015. The diversity and host interactions of *Propionibacterium acnes* bacteriophages on human skin. *ISME Journal* **9(9):**2116 DOI 10.1038/ismej.2015.47.

Ly M, Abeles SR, Boehm TK, Robles-Sikisaka R, Naidu M, Santiago-Rodriguez T, Pride DT. 2014. Altered oral viral ecology in association with periodontal disease. *mBio* **5(3):**e01133-14 DOI 10.1128/mBio.01133-14.

Ma Y, Madupu R, Karaoz U, Nossa CW, Yang L, Yooseph S, Yachimski PS, Brodie EL, Nelson KE, Pei Z. 2014. Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. *Journal of Virology* **88(9):**4786–4797 DOI 10.1128/JVI.00093-14.

Malim MH, Emerman M. 2001. HIV-1 sequence variation: drift, shift, and attenuation. *Cell* **104(4):**469–472 DOI 10.1016/s0092-8674(01)00234-3.

Marinelli LJ, Fitz-Gibbon S, Hayes C, Bowman C, Inkeles M, Loncaric A, Russell DA, Jacobs-Sera D, Cokus S, Pellegrini M, Kim J, Miller JF, Hatfull GF, Modlin RL. 2012. *Propionibacterium acnes* bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *mBio* **3(5):**e00279-12 DOI 10.1128/mBio.00279-12.

Meisel JS, Hannigan GD, Tyldsley AS, SanMiguel AJ, Hodkinson BP, Zheng Q, Grice EA. 2016. Skin microbiome surveys are strongly influenced by experimental design. *Journal of Investigative Dermatology* **136(5):**947–956 DOI 10.1016/j.jid.2016.01.016.

Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America* **110(30):**12450–12455 DOI 10.1073/pnas.1300833110.

Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD. 2012. Hypervariable loci in the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America* **109(10):**3962–3966 DOI 10.1073/pnas.1119061109.

Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research* **21(10):**1616–1625 DOI 10.1101/gr.122705.111.

Mistry N, Wibom C, Evander M. 2008. Cutaneous and mucosal human papillomaviruses differ in net surface charge, potential impact on tropism. *Virology Journal* **5(1):**118 DOI 10.1186/1743-422x-5-118.

Nishida K, Frith MC, Nakai K. 2009. Pseudocounts for transcription factor binding sites. *Nucleic Acids Research* **37(3):**939–944 DOI 10.1093/nar/gkn1019.

Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160(3):**447–460 DOI 10.1016/j.cell.2015.01.002.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9(1):**312 DOI 10.1186/1471-2164-9-312.

Nuxoll AS, Halouska SM, Sadykov MR, Hanke ML, Bayles KW, Kielian T, Powers R, Fey PD. 2012. CcpA regulates arginine biosynthesis in *Staphylococcus aureus* through repression of proline catabolism. *PLoS Pathogens* **8(11):**e1003033 DOI 10.1371/journal.ppat.1003033.

Oh J, Byrd AL, Deming C, Conlan S, Program NCS, Kong HH, Segre JA. 2014. Biogeography and individuality shape function in the human skin metagenome. *Nature* **514(7520):**59–64 DOI 10.1038/nature13786.

Quint KD, Genders RE, de Koning MN, Borgogna C, Gariglio M, Bouwes Bavinck JN, Doorbar J, Feltkamp MC. 2015. Human Beta-papillomavirus infection and keratinocyte carcinomas. *Journal of Pathology* **235(2):**342–354 DOI 10.1002/path.4425.

Rambaut A. 2006. FigTree. *Available at http://tree.bio.ed.ac.uk/software/figtree/* (accessed 3 June 2015).

Schiller JT, Lowy DR. 2012. Understanding and learning from the success of prophylactic human papillomavirus vaccines. *Nature Reviews Microbiology* **10**:681–692 DOI 10.1038/nrmicro2872.

Schillinger T, Lisfi M, Chi J, Cullum J, Zingler N. 2012. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics* **13**:430 DOI 10.1186/1471-2164-13-430.

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P. 2013. Genomic variation landscape of the human gut microbiome. *Nature* **493(7430):**45–50 DOI 10.1038/nature11711.

Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9(1):**34 DOI 10.1186/1471-2105-9-34.

Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6(3):**e17288 DOI 10.1371/journal.pone.0017288.

Shah S, Alexaki A, Pirrone V, Dahiya S, Nonnemacher MR, Wigdahl B. 2014. Functional properties of the HIV-1 long terminal repeat containing single-nucleotide polymorphisms in Sp site III and CCAAT/enhancer binding protein site I. *Virology Journal* **11(1):**92 DOI 10.1186/1743-422x-11-92.

Shin J-H, Blay S, McNeney B, Graham J. 2006. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software* **16**: DOI 10.18637/jss.v016.c03.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9):**1312–1313 DOI 10.1093/bioinformatics/btu033.

UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **42(D1):**D191–D198 DOI 10.1093/nar/gkt1140.

Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA. 2013. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research* **41(D1):**D571–D578 DOI 10.1093/nar/gks984.

Varga M, Kuntová L, Pantůček R, Mašlaňová I, Růžičková V, Doškař J. 2012. Efficient transfer of antibiotic resistance plasmids by transduction within methicillin-resistant *Staphylococcus aureus* USA300 clone. *FEMS Microbiology Letters* **332(2):**146–152 DOI 10.1111/j.1574-6968.2012.02589.x.

Vinzón SE, Braspenning-Wesch I, Müller M, Geissler EK, Nindl I, Gröne H-J, Schäfer K, Rösl F. 2014. Protective vaccination against papillomavirus-induced skin tumors under immunocompetent and immunosuppressive conditions: a preclinical study using a natural outbred animal model. *PLoS Pathogens* **10(2):**e1003924 DOI 10.1371/journal.ppat.1003924.

Wang J, Aldabagh B, Yu J, Arron ST. 2014. Role of human papillomavirus in cutaneous squamous cell carcinoma: a meta-analysis. *Journal of the American Academy of Dermatology* **70(4):**621–629 DOI 10.1016/j.jaad.2014.01.857.

Warnes G, Gorjanc G, Leisch F, Man M. 2003. Genetics: Population Genetics. *Available at https://CRAN.R-project.org/package=genetics* (accessed 20 May 2016).

Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. 2006. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct* **1**:34 DOI 10.1186/1745-6150-1-34.

Yates CM, Filippis I, Kelley LA, Sternberg MJE. 2014. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology* **426(14):**2692–2701 DOI 10.1016/j.jmb.2014.04.026.

Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, Cao K, Wang L-S, Gregory BD. 2010. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. *PLoS Genetics* **6(9):**e1001141 DOI 10.1371/journal.pgen.1001141.