 doi:10.20944/preprints202108.0149.v1

Article

HPV DeepSeq: Ultra-fast method of NGS data analysis and visualization using automated workflows and a customized Papillomavirus database in CLC Genomics Workbench

Jane Shen-Gunther [1,*], Qingqing Xia [2], Hong Cai [3,4] and Yufeng Wang [3,4,*]

[1] Gynecologic Oncology & Clinical Investigation, Department of Clinical Investigation, Brooke Army Medical Center, Fort Sam Houston, TX 78234, USA
[2] Department of Clinical Investigation, Brooke Army Medical Center, Fort Sam Houston, TX 78234, USA
[3] Department of Biology, University of Texas at San Antonio, San Antonio, TX 78249, USA
[4] South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX 78249
* Correspondence: jane.shengunther.mil@mail.mil or shengunther@livemail.uthscsa.edu

Abstract: Next-generation sequencing (NGS) has actualized human papillomavirus (HPV) virome profiling for in-depth investigation of viral evolution and pathogenesis. However, viral computational analysis remains a bottleneck due to semantic discrepancies between computational tools and curated reference genomes. To address this, we developed and tested automated workflows for HPV taxonomic profiling and visualization using a customized Papillomavirus database in CLC Microbial Genomics Module. HPV genomes from Papilloma Virus Episteme were customized and incorporated into CLC "ready-to-use" workflows for stepwise data processing to include: 1) Taxonomic Analysis, 2) Estimate Alpha/Beta Diversities, and 3) Map Reads to Reference. Low-grade (n = 95) and high-grade (n = 60) Pap smears were tested with ensuing collective runtimes: Taxonomic Analysis (36 min); Alpha/Beta Diversities (5 sec); Map Reads (45 min). Tabular output conversion to visualizations entailed 1-2 keystrokes. Biodiversity analysis between low- (LSIL) and high-grade squamous intraepithelial lesions (HSIL) revealed loss of species richness and gain of dominance by HPV-16 in HSIL. Integrating clinically relevant, taxonomized HPV reference genomes within automated workflows proved to be an ultra-fast method of virome profiling. The entire process named "HPV DeepSeq" provides a simple, accurate and practical means of NGS data analysis for a broad range of applications in viral research.

Keywords: Bioinformatics; Cervical cancer; Deep Sequencing; Human papillomavirus; HPV genotyping; Metagenome; Next generation sequencing; Taxonomic classification; Virome

1. Introduction

Hippocrates was the first to describe cervical cancer and its destructive nature around 400 BCE [1]. Two thousand years elapsed before zur Hausen and his "papillomavirus crew" made the breakthrough discovery of identifying human papillomavirus (HPV)-16 in cervical, vulvar, and penile cancers in 1983 [1–3]. Since then, the causal role of carcinogenic HPV in anogenital, oropharyngeal, and dermatological cancers (in patients with *epidermodysplasia verruciformis*) has been established and classified by the International Agency for Research on Cancer (IARC) [4,5]. The etiological role of HPV in breast and esophageal cancers has been postulated for several decades but remains controversial due to conflicting findings [6,7]. Nonetheless, HPV is the second most prevalent primary infectious cause of cancer worldwide. The annual global burden of 570,000 new cervical and 120,000 other anogenital and oropharyngeal cancer cases have been attributed to HPV [8,9].

The Papillomavirus (PV) is a small 8,000 base pair (bp) double-stranded, circular DNA virus that co-evolved with an ancestral host over 400 million years [10]. The PV

genome backbone acquired oncogenes E6 and E7 and later E5 approximately 184 and 55 million years ago, respectively [10]. The genetic differences in these oncogenes conferred disparate phenotypes and oncogenic potential [10,11]. Recent phylogenetic analysis also suggested that anatomical site predilection and tissue tropism by distinct HPV genotypes may be a result of viral niche adaptation to host ecosystems [12]. Site-specific genotypes and virome composition may be further shaped by the host's immune response [13]. Therefore, anatomical virome characterization is crucial to our understanding of niche-specific, virus-host adaptive evolution foundational to pathogenesis.

Recent advancements in next-generation sequencing (NGS) have actualized HPV virome profiling [14]. Since the commercialization of high-throughput sequencing (HTS) instruments in 2005, the variety and choices of sequencing technologies, chemistries, platforms, capacities, and kits have expanded exponentially [15,16]. The wet lab portion of genomics research has become more streamlined, simpler, and accessible to the researcher. However, bioinformatics analysis remains a bottleneck [17,18]. First, the enormous amounts of complex data generated from each sample is non-trivial. This is compounded by disparate, open-source tools which are often command-line based and require advanced computational or coding skills [17,18]. Second, viral taxonomies based on the International Committee on Taxonomy of Viruses (ICTV) are subject to new revisions. In 2019, the ICTV taxonomic structure was revised from a 5-rank (1991-2017) to a 15-rank structure which imposes retooling of existing software and restructuring of reference database(s) for viral metagenomic analysis [19]. Furthermore, the 15-rank format is inconsistent with that of open-source Quantitative Insights into Microbial Ecology (QIIME) popularly used for bacterial and fungal taxonomic analysis [20]. To circumvent these two daunting barriers (i.e., taxonomic structure and software), we customized the curated Papilloma Virus Episteme (PaVE) database [21] for use in a user-friendly, GUI-based, commercial software for sequence analysis. We aimed to develop and test automated workflows for HPV taxonomic profiling and visualization within the CLC Microbial Genomics Module (MGM) plugged into the main Workbench (WB). A simple, rapid, and accurate means of NGS data analysis will ultimately propel HPV research and serve a broad range of applications from discovery to therapeutics.

2. Results

2.1. Taxonomic classification and visualization of HPV metagenomes

This dataset included 155 cytology samples, classified as low-grade squamous intraepithelial lesions (LSIL) ($n = 95$) and high-grade squamous intraepithelial lesions (HSIL) ($n = 60$). The "Data quality control (QC) and taxonomic profiling" workflow computational runtime on this dataset was 23:33 and 12:50 minutes for the LSIL and HSIL samples, respectively (Table 1). The QC workflow generated the following outputs: 1) QC for sequencing reads (graphical report and supplementary report) and 2) Abundance table. Specifically, the graphical report summarized the total number of sequences and nucleotides in a sample, per-sequence analysis, per-base analysis, over-representation analyses, sequence duplication levels, and duplicated sequences. The QC supplementary report includes two additional columns (i.e., "coverage" and "abs") for absolute numbers of sequences or bases for the per-sequence or per-base analyses. An in-depth explanation of the QC metrics is beyond the scope of this article. The reader is referred to the CLC MGM manual online for details.

 doi:10.20944/preprints202108.0149.v1

Table 1. Software performance efficiency.

Workflow/Tool[1]	Input file			Output	Runtime[2]	
	n	type	size	report/table	total	unit
Data QC & Taxonomic profiling	155	.clc[3]	10.6 GB	QC graphical reports Abundance table[4]	00:36:23	00:00:14
Alpha and Beta diversities	1	.clc[5]	4 KB	Diversity plots Distance matrix table[4]	00:00:05	<00:00:01
Map reads to reference	155	.clc[3]	10.6 GB	Mapping report Reads track	00:45:24	00:00:18
Differential abundance analysis	1	.clc[5]	4 KB	Experiment table Statistical result table	00:00:03	<00:00:01
Convert abundance table to exp	1	.clc[5]	4 KB	Experiment table Statistical result table	00:00:02	<00:00:01
Create heat map for abun table	1	.clc[5]	4 KB	Heat map chart	00:00:01	<00:00:01
BLAST[5]	155	.phd	947 KB	BLAST table	00:00:12	<00:00:01

abun; abundance; BLAST, Basic Local Alignment Search Tool; .clc, CLC file format; dist., distance; E6/E7, HPV E6/E7 gene amplified by PCR; exp, experiment; HSIL, high-grade squamous intraepithelial lesion; QC, quality control; sec, second. [1] All workflows and tools were tested with the full dataset (*n* = 155) of samples. CLC pre-built workflows tested: 1) Data QC & Taxonomic profiling with integrated HPV reference genome database, 2) Alpha and Beta diversities, and 3) Map reads to reference. CLC microbial genomics tools tested: a) Differential abundance analysis, b) Convert abundance table to experiment, c) Create heat map for abundance table, and d) BLAST with HPV BLAST database. [2] Time notation (hh:mm:ss) represents the number of complete hours (hh), minutes (mm) and seconds (ss). Total time is the total runtime for 155 samples. Unit time is the mean runtime per sample. [3] Paired fastq sequences in ".clc" format. [4] Table may be visualized as a chart with 1-click. [5] Merged LSIL and HSIL abundance table with appended metadata in ".clc" format. [6] BLAST (*blastn* program) united with the HPV reference and variant genome database was tested for genotyping Sanger sequences. Total runtime comprised of importing 155 sequences (1 sec) + BLAST (11 sec).

The taxonomic profiling workflow generated individual abundance tables that display the names of the identified taxa, 7-level taxonomic nomenclature, coverage estimate, and abundance value (raw or relative number of reads found in the sample associated with the taxon). A merged abundance table inclusive of multiple samples also displays summary statistics (e.g., combined abundance of reads for the taxon across all samples, and the minimum, maximum, mean, median and standard deviation of the number of reads for the taxa across all samples) (Supplementary Table 1). Additional reports, such as, "Reads (mapping to database or host)" and "Reads (umapped)" may also be selected for output. However, this is performed at the cost of increased runtime.

The abundance table may be visualized as a stacked bar chart or sunburst plot with 1-click (Figure 1A-C). Metadata added and used to aggregate groups of samples (i.e., cytological grades "LSIL" and "HSIL") produced the 2-group stacked bar chart for comparison of HPV genotype composition (Figure 1B). The sunburst plots display hierarchical taxonomic ranks revealing distinct differences in the HPV communities between LSIL and HSIL groups (Figure 1C).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

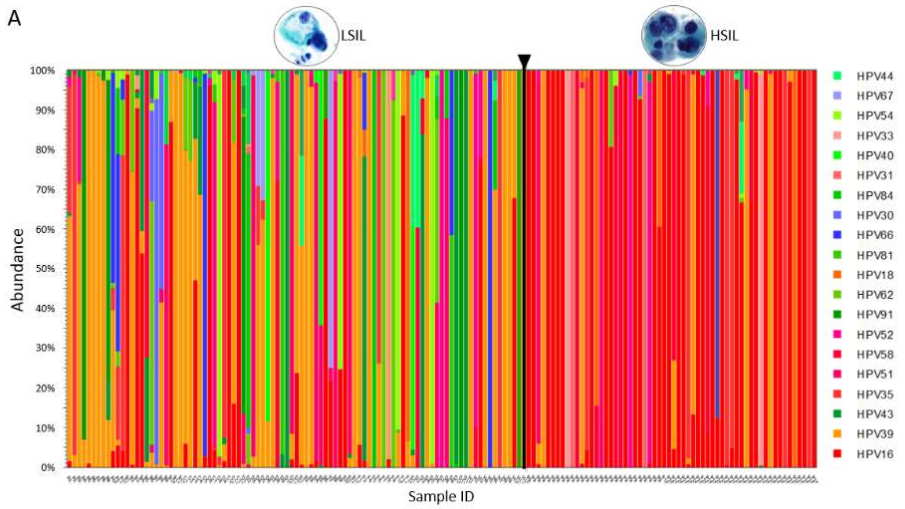


Figure 1. Taxonomic profiling results based on the HPV Reference Index. (**A**) Abundance of HPV genotypes found in individual LSIL and HSIL samples (stacked bars). Deep sequencing of HPV E6/E7 amplicons derived from each LSIL (n = 95) or HSIL (n = 60) sample identified 32 unique HPV genotypes with the top 20 shown (legend) and quantitated their composition (%) based on abundance (n) of mapped reads to total mapped reads. (**B**) HPV genotype composition of samples grouped by cytological grade i.e., LSIL and HSIL. Visualization and comparison of grouped samples (stacked bars) revealed the dominant genotypes in LSIL and HSIL as HPV-39 (46%) and HPV-16 (69%), respectively with significant changes in proportional composition (Baggerley's test, *p-value (Bonferroni) < 0.001). (**C**) Sunburst plots visualize hierarchical data outwardly from parent to child nodes. Here, sunburst plots reveal distinct differences in the HPV communities according to seven taxonomic ranks, specifically, the last two ranks (genus/species and genotypes) between LSIL and HSIL.

Both Sanger and NGS sequencing were used to detect HPV genotypes and sub-lineages within each sample. Sanger sequencing resolved the single dominant HPV genotype within each sample. Compared to Sanger sequencing, NGS achieved a higher resolution in detection of mixed genotypes (cut-off at six genotypes) and low-abundance genotypes (cut-off at ≥ 1% of total composition) (Supplementary Table 2). Comparing the dominant genotypes and sub-lineages (variants) derived from both sequencing methods, the inter-assay reliability was near-perfect (kappa = 0.94) (Table 2). Samples (9/155 samples, 5.81%) with discordant HPV genotyping results may be explained by: 1) low-quality Sanger sequence (BLAST "not available") (n = 1), 2) low-quality Sanger sequence (BLAST max id < 90% and max bit score < 531) (n = 2), 3) L1 and E6/E7 primer bias (n = 3), or 4) E6 sequencing primer bias (n = 3). The complete Sanger/BLAST table with 9 discordant results (bold) are shown in Supplementary Table 3.

Table 2. HPV genotype concordance: Sanger seq/BLAST vs. Deep seq/Taxonomic profiling.

Agreement Statistic	LSIL	HSIL	LSIL/HSIL
Samples[1] (n)	95	60	155
Discordant[2] (n, %)	4 (4.21%)	5 (8.33%)	9 (5.81%)
Agreement[3] (n, %)	91 (95.79%)	55 (93.22%)	146 (94.81%)
Expected Agreement	12.95%	34.50%	15.11%
Kappa	0.9516	0.8965	0.9388
Std. Error	0.0345	0.0613	0.03
p-value	<0.0001	<0.0001	<0.0001

BLAST, Basic Local Alignment Search Tool; E6/E7, HPV E6/E7 gene amplified by PCR; HSIL, high-grade squamous intraepithelial lesion; HPV, human papillomavirus; L1, HPV L1 gene amplified by PCR; LSIL, low-grade squamous intraepithelial lesion; seq, sequencing. [1] Samples with HPV

 doi:10.20944/preprints202108.0149.v1

genotype(s) determined concurrently by Sanger sequencing/BLAST and deep sequencing/taxonomic profiling from HPV E6/E7 amplicons. The top ranking (most abundant and qualified) HPV genotype identified by taxonomic profiling was compared to the highest bit scoring (best) BLAST genotype. For non-sequenceable or interpretable HPV E6/E7 Sanger results, HPV L1 Sanger results were used alternatively. Detailed sequencing and genotyping results for individual samples (n = 155) are provided in Supplementary Table 2. [2] Samples with discordant HPV genotype results as determined by taxonomic profiling and BLAST. [3] HPV genotype agreement between the two sequencing/genotyping methods. For HPV variant lineage agreement, the result was nearly 100% agreement except for one sample (PC_2675) where the variant lineage of HPV type 68 could not be determined by deep sequencing. Distinct genotypes, variants, and sub-lineages, by definition, have >10%, 1.0%, and 0.5% to 1.0% nucleotide sequence difference, respectively [21].

By convention, sequences must possess ≥ 90% homology with the closest known type for HPV genotyping assignment [21]. Correlation analysis was performed to find the corresponding BLAST max bit score and E-value to BLAST 90% maximum identity. Two relationships emerged: 1) BLAST maximum bit score and $\log_{10}$ (E-value) were perfectly, linearly anti-correlated, and 2) BLAST max bit score and maximum identity (%) were curvilinearly correlated (Supplementary Figure 1). Taken together, BLAST maximum identity of ≥ 90% corresponded to a maximum bit score of ≥ 531 (equivalent to E-value ≤ $\log_{10}$-150) which were used as the threshold for quality genotyping results. BLAST values less than this threshold were considered as "uncertain" HPV genotyping result.

2.2. Diversity analysis and visualization of LSIL/HSIL HPV communities

The "Merge and Estimate Alpha and Beta Diversities" workflow runtime on this dataset was 5 seconds (Table 1). The diversity analyses workflow generated the following outputs: 1) alpha diversity rarefaction table and plots, and 2) beta diversity distance matrix and principal coordinate analysis (PCoA) plots.

The alpha diversity metric is calculated by sub-sampling the abundances at different depths (number of reads) in each sample. The rarefaction analysis parameters define the granularity of the alpha diversity curve as presented in Figure 2. Abundance table metadata used for aggregating groups of samples (i.e., "LSIL" and "HSIL") produced the box plot with auto-calculated Mann-Whitney U statistical results in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

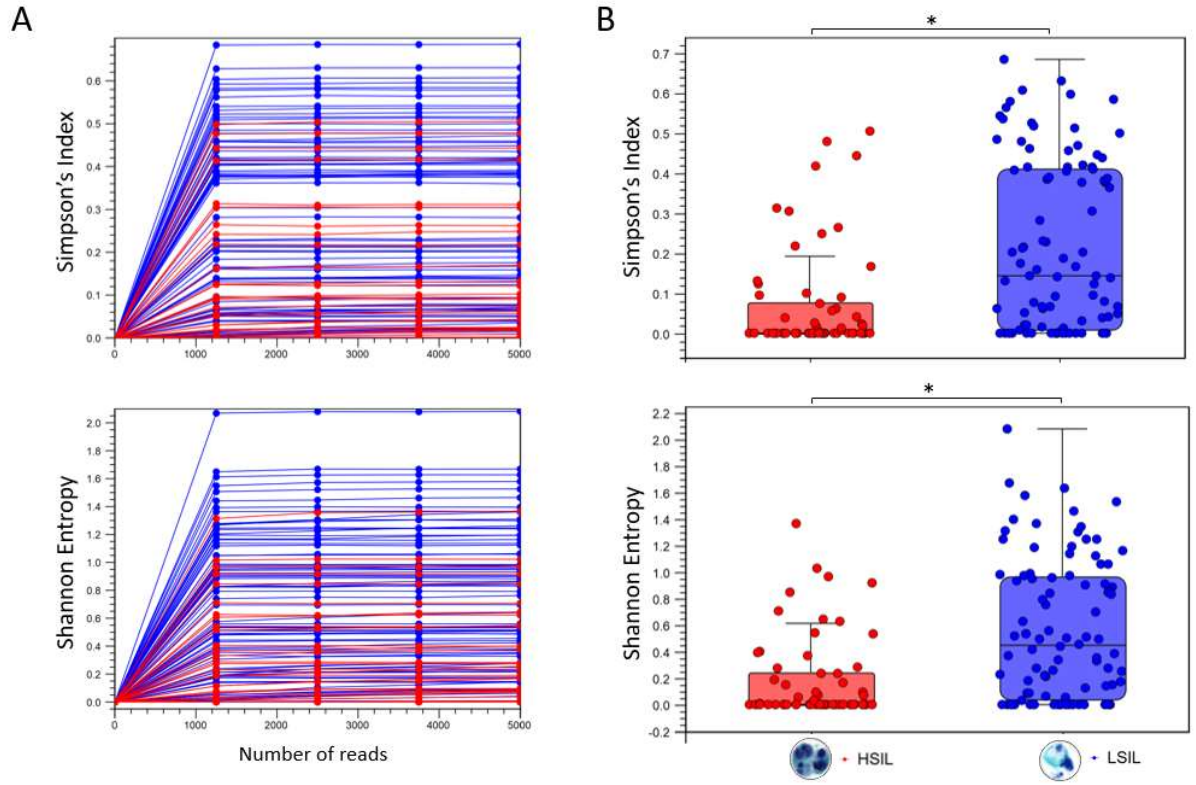


Figure 2. Diversity analysis of HPV genotypes in LSIL and HSIL samples. (**A**) Alpha diversity rarefaction curves estimate the indices for HPV genotypes in LSIL (n = 95) and HSIL (n = 60) samples. (**B**) Summary statistics are shown as boxplots after categorical grouping. A total of 29 unique genotypes were found in LSIL versus 22 genotypes for HSIL. Species richness measured by Simpson's index showed a reduction from LSIL (median = 0.146) to HSIL (median = 0.062) samples (Mann-Whitney test, *p < 0.001). The respective Shannon indices for LSIL (median = 0.454) and HSIL (median = 0.215) were indicative of reduced diversity with disease progression (Mann-Whitney test, *p < 0.001).

The workflow output for beta diversity analysis was a Bray-Curtis distance matrix between samples and PCoA plots in 2D or 3D. The 3D PCoA plot visually displayed the dissimilarities in HPV composition between all samples (Figure 3A). After grouping HSIL and LSIL samples, the dissimilar HPV communities and the most influential genotypes were visually apparent (Figure 3B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

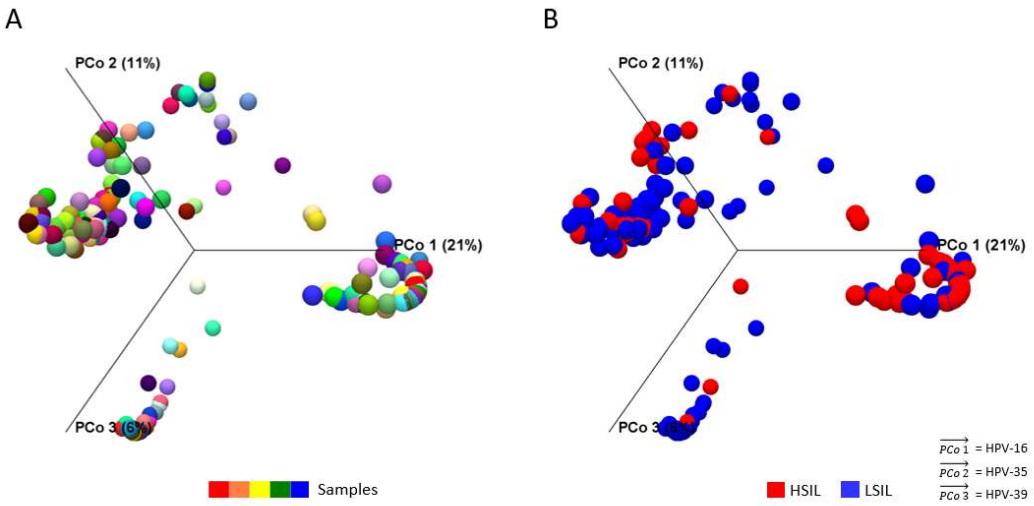


Figure 3. HPV community structures between LSIL and HSIL. 3D-Principal Coordinate Analysis (PCoA) plots of samples before (**A**) and after (**B**) grouping by cytological category. After grouping by LSIL and HSIL, HPV-16, -35, -39 were identified as the three most influential genotypes (vectors) in both HPV communities. Dissimilarity between the two HPV communities was HPV-16 (PCoA 1) being the most influential genotype in HSIL versus HPV-39 for LSIL (PCoA 3). β-diversity was measured by Bray-Curtis index (PERMANOVA, $*p < 0.05$).

2.3. Differential abundance analysis and visualization of LSIL/HSIL HPV communities

The "Convert Abundance Table to Experiment" and "Proportion-based Statistical Analysis" runtimes on this dataset were 2 seconds and <1 second, respectively (Table 1). The output generated a table listing the differential abundance and weighted proportions difference of each HPV type between LSIL and HSIL groups concomitant with statistical analysis (Baggerley's test-statistic, p-value; Bonferroni and FDR corrected p-values).

The "Create Heat Map for Abundance Table" runtime on this dataset was <1 second. The resultant heat maps revealed hierarchal clustering of HPV genotypes on the 155 LSIL/HSIL samples. HSIL and LSIL samples with HPV genotypes, such as, HPV-16 and -39 with similarly high abundances (~100% of reads) are noted in the blue-red spectrum versus dissimilar abundance in the red spectrum (Figure 4A). The aggregated heat map of LSIL and HSIL samples revealed dissimilar HPV profiles with the prevailing genotypes shown in red (Figure 4B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

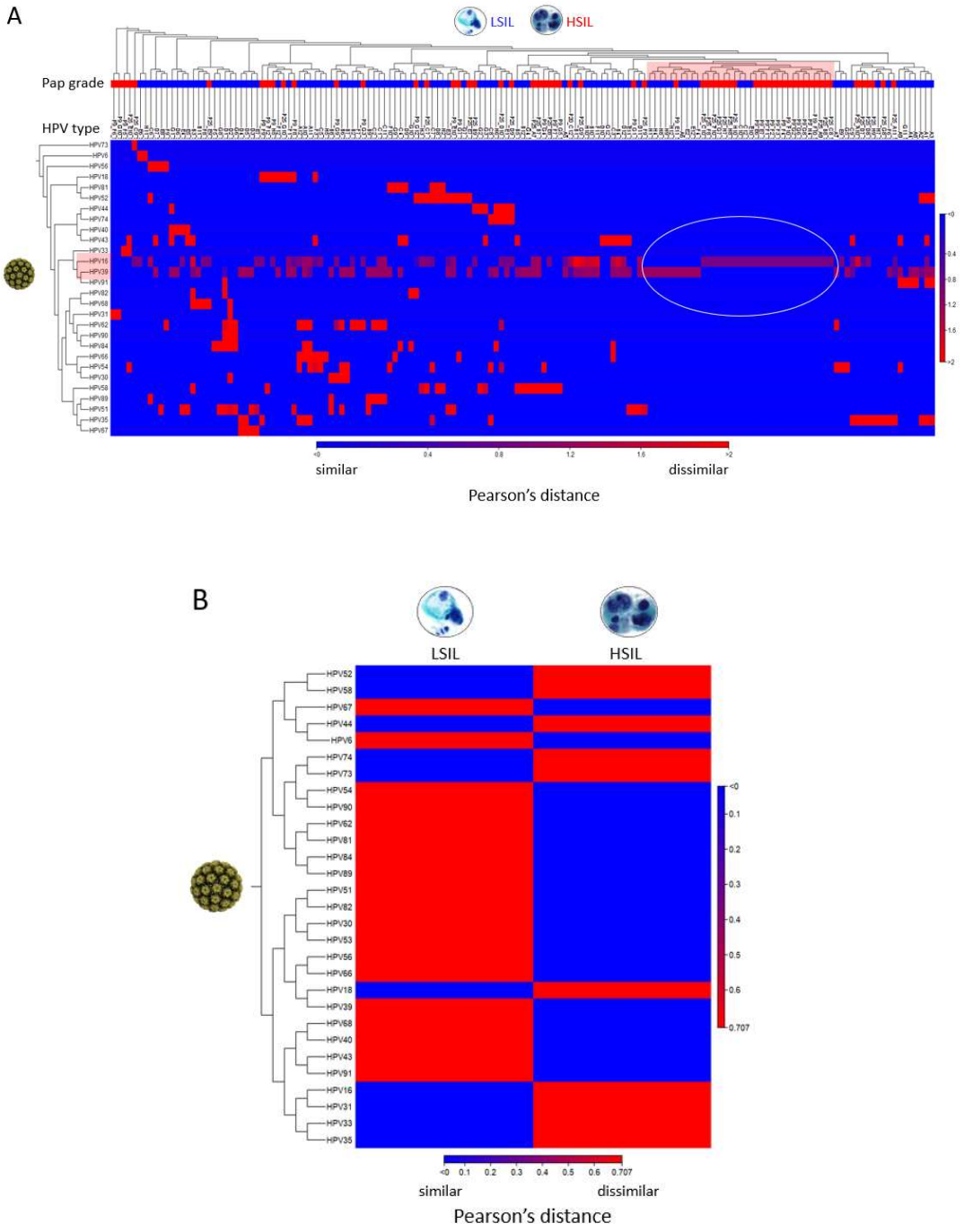
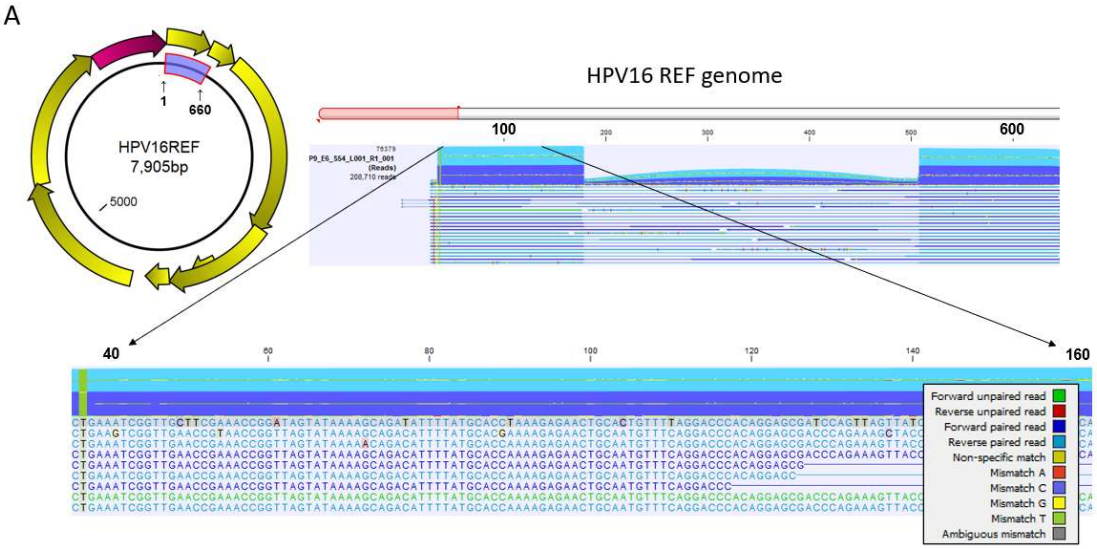


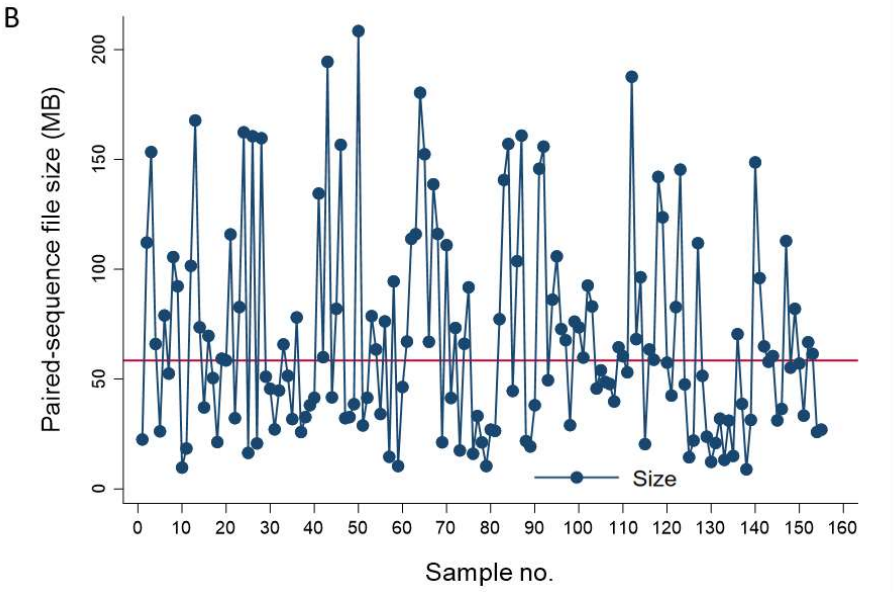
Figure 4. Clustered heat map of HPV genotypes among LSIL/HSIL samples. (**A**) The heat map represents hierarchal clustering of HPV genotypes on LSIL/HSIL samples (n = 155) and in reverse along the Y and X axes, respectively. The dissimilarity measure i.e., Pearson's distance (1- | Pearson correlation|) quantifies the dissimilarity in the variables of interest i.e., HPV genotypes between individual samples. The agglomerative clustering method identifies the originating, most similar pair of clusters (pink shade). In this dataset, single (pure) infections of HPV-16 in HSIL and HPV-39 in LSIL (encircled) are the closest clusters. From this point, cluster divergence toward the left part of the heat map reveals increasingly, heterogeneous HPV infections predominantly in LSIL samples. (**B**) Aggregated heat map of LSIL (n = 95) and HSIL (n = 60) samples reveal dissimilar HPV profiles useful for data simplification and taxonomy development. The prevailing genotypes in LSIL or HSIL are shown in red. The color scales show the spectrum of Pearson's distance between 0 (blue) and 2 (red) indicating similar and dissimilar correlation coefficients, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

2.4. Read mapping and visualization of mapped tracks

The "Map Reads to Reference" workflow runtime on this dataset (n = 155) was 45:24 minutes with a mean of 18 seconds per sample (Table 1). The workflow generated two outputs: 1) mapping report and 2) reads track. Specifically, the mapping report summarized the total number of reads, mapped/unmapped reads, intact/broken paired reads, and matched/unmatched read length distribution per sample. A representative reads track shows paired-reads of a sample mapped onto the linearized HPV-16 reference genome (Figure 5A). Zooming in allowed visualization of the sequences down to the nucleotide level for comparison to the reference genome and detection of variants. Detailed analysis of mapping time revealed a high correlation between runtime and number of merged sequences in a sample. Furthermore, the number of merged sequences showed a linear correlation with input file size (MB). These correlative relationships are shown in Figure 5B-D. The fitted lines may be used independently or jointly to estimate mapping runtimes from sample file size or number of merged sequences.





Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

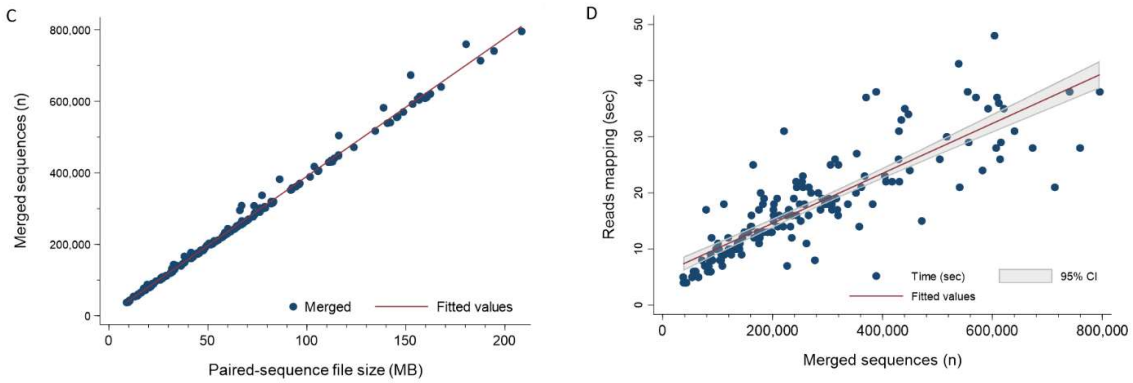


Figure 5. Read Mapping. (**A**) The HPV E6/E7 (660bp) gene segment highlighted in blue on the circular prototypical HPV-16 genome (GenBank ID: K02718) is the target used for amplicon sequencing and genotyping. The Map Reads to Reference workflow output displays the reads mapped on to the linearized HPV reference genome. Zooming in from the whole genome window (top) allows viewing of the sequences down to the nucleotide level (bottom). The color-coding legend defines the corresponding read types and nucleotide mismatches. (**B**) NGS paired-sequence file size for each of the 155 study samples. The connected scatterplot reveals the extent of file size variation between samples. The median (—) was 58.5 MB (range, 8.9-208.5). (**C**) Scatterplot between NGS paired-sequence file size (MB) and merged sequences (*n*) for the 155 study samples showed near-perfect linear correlation (R^2 = 0.9945). The regression line (merged sequences = 3,415 + 3,868 * file size) is shown as (—). (**D**) Merged sequences (*n*) and reads mapping time (s) for the study cohort (*n* = 155) were highly correlated (R^2 = 0.7233) as shown by the scatterplot and regression line (mapping time = 5.7 + 4.44E-5 * merged sequences) (—). The equations above may be used jointly or independently to estimate total mapping time based on file size or number of sequences.

3. Discussion

In this study, we developed and tested several workflows and tools for HPV virome profiling from deep sequenced clinical samples. By integrating our taxonomically customized HPV genome database within CLC MGM workflows, we were able to traverse disparate computational processes using one multi-functional software. Taxonomic classification and visualization of HPV metagenomes were accomplished efficiently and quickly to reveal differences between LSIL and HSIL viral communities. Alpha and Beta diversity analyses were processed in <5 seconds to quantify the loss of α-diversity and gain of dominance by HPV-16 in HSIL over HPV-39 in LSIL samples. Similarly, differential abundance analysis and heat map visualization of LSIL/HSIL HPV communities were achieved within seconds to reveal dissimilar HPV profiles. The "Map Reads to Reference" workflow consumed more time than "Taxonomic Profiling" due to inherent computational complexity. The processing time correlated linearly with the number of merged sequences within a sample. The resulting mapped tracks with zoomable visualization provided easy inspection of mapped regions and detection of variants at the nucleotide-level. Finally, HPV genotyping results by NGS/taxonomic profiling was corroborated by concurrent Sanger/BLAST. In fact, NGS provided a much richer picture of the virome and evolutionary dynamics between disease states (i.e., LSIL to HSIL) than ever possible with conventional sequencing.

The strength of the methods developed herein is the integration of a taxonomized database with automated workflows for viral metagenomic analysis. The desired end state for the software user is a single, user-friendly, multi-functional tool, analogous to the "Swiss Army knife" [22]. We also strived to achieve the most efficient data processing pipeline by minimizing steps and time. Furthermore, testing the workflows on a portable notebook computer endorses software performance efficiency for benchtop or fieldwork, as well as, in austere clinical settings. In contradistinction, a recent publication used nine bioinformatics software for HPV de novo assembly, genotyping, phylogenetic analysis,

 doi:10.20944/preprints202108.0149.v1

and visualization (Bowtie2, SPAdes, VAPiD, MAFFT, RAxML, MEGA7, iTOL5.3, Dendroscope3, and R) [23]. In recent years, the list of superb bioinformatics software has lengthened considerably and is welcomed by the research community [18,24]. However, for the clinical virologist or physician-scientist, it is time-consuming and impractical to learn these sophisticated programs. Instead, a single, integrated, easy-to-use platform is preferred.

Another strength of this study is the corroboration of NGS/taxonomic profiling results by Sanger/BLAST. For the 9/155 (5.81%) discordant results, the quality of Sanger sequences was suboptimal in 3 samples, and primer or sequencing bias was the probable cause in 6 others. By convention, NGS results necessitate validation by Sanger sequencing as the reference standard. However, recently the necessity of Sanger validation has been called into question for human genome sequencing by several studies [25,26]. Arteche-Lopez et al., reported their validation study of 1,109 NGS variants in 825 clinical exomes, the largest sample set to date using Illumina technology [25]. Only 3 discrepancies were found, and all false negative results arose from Sanger sequencing [25]. Taken together, high-quality NGS and analytical methods offer higher resolution and accuracy than Sanger/BLAST and only selective validation may be necessary. We acknowledge that our study has limitations that is, other cytological categories with potentially different HPV virome profiles were not included for comparison. To bridge this gap, we intend to analyze our ongoing large-scale study (>3,000 samples) with the same workflows and streamline comparative analysis of three or more cytological categories.

4. Materials and Methods

4.1. Subjects, samples, and deep sequencing

A subset of HPV-positive LSIL (n = 95) and HSIL (n = 60) cervical cytology samples were randomly selected from a larger (3,000+ samples) Congressionally Directed Medical Research Programs (CDMRP) grant-funded project to test our workflow-based methods described herein. The residual liquid-based cervical cytology samples were procured consecutively from the Department of Pathology after completion of cytological diagnosis. Demographic and cytological data were abstracted from the electronic health record (AHLTA) as metadata for association with taxonomic profiling results.

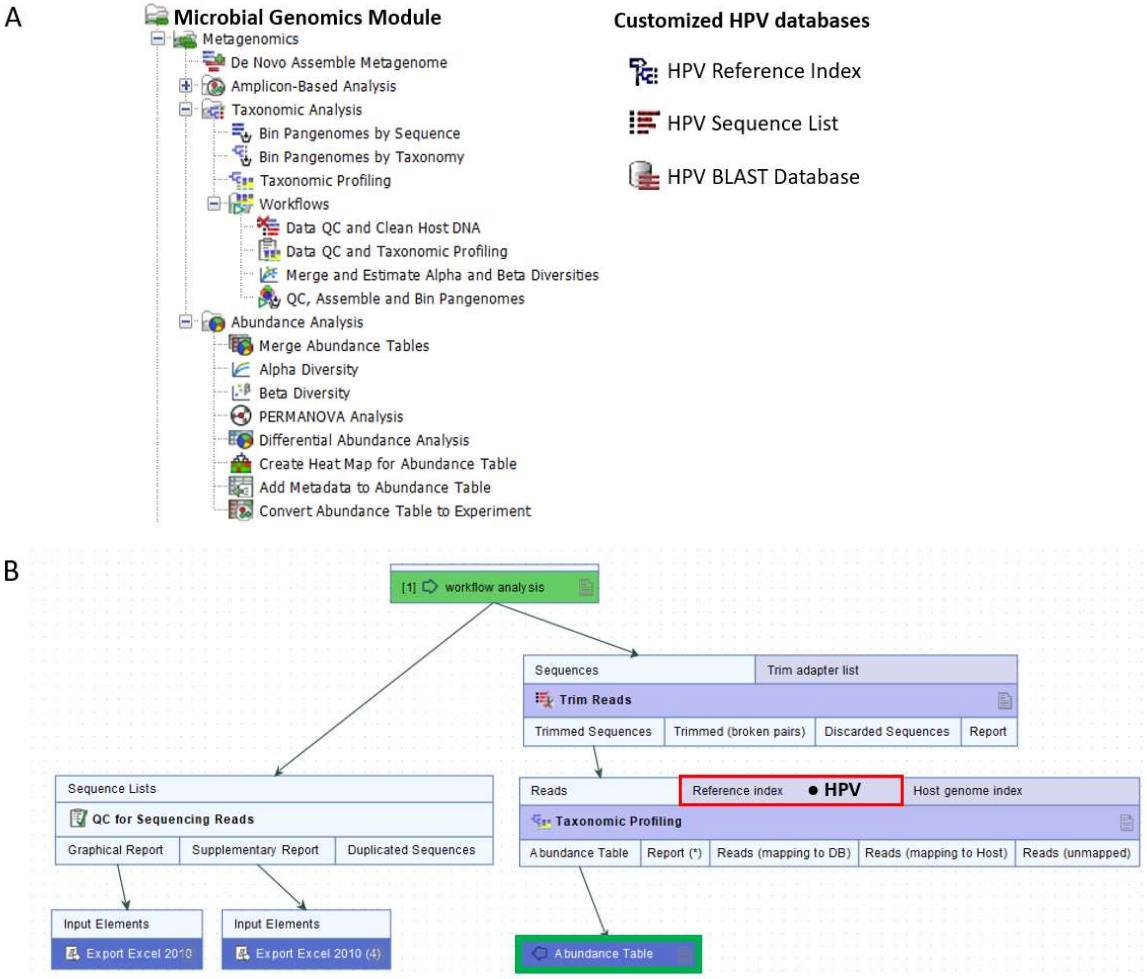
DNA extraction, HPV DNA amplification and deep sequencing were performed as described previously with a few modifications below [14,27]. Cellular DNA extraction was performed after off-board cellular lysis using the QIAsymphony DSP DNA Midi Kit (96) in a QIAsymphony robotic workstation (Qiagen). HPV DNA was amplified as previously published using consensus primers to target a 660 bp region of the E6/E7 gene (nucleotide positions 28 to 658 on the HPV-16 genome) for genotyping [14,27]. The PCR products were purified using PureLink (ThermoFisher Scientific) spin column-based method and sent to Lucigen (Middleton, WI) for next-generation sequencing. The submitted E6/E7 amplicons were mechanically sheared to 300-500 bp prior to construction of DNA libraries using the NxSeq AmpFREE low DNA library kit (Lucigen) per protocol. The four primary steps in library construction were: 1) end repair, 2) a-tailing, 3) adaptor ligation, and 4) size selection. The libraries were normalized quantitatively for equimolar representation from each sample prior to pooling and sequencing. Paired-end bi-directional sequencing (2 × 150 bp) was performed on the MiSeq (Illumina) instrument using the MiSeq Reagent Kit v2 (300-cycles) for bridge amplification.

The PCR products were concurrently subjected to dideoxy (Sanger) sequencing for validation of deep-sequenced results. Briefly, amplicons (~200 ng DNA/sample) were sequenced using primer GP-E6-3F at Eurofins Genomics (Louisville, KY). The resulting sequences were BLAST aligned for HPV genotyping as described above.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

4.2. Customized HPV reference databases for CLC workflows

HPV reference (n = 219) and variant genomes (n = 136) from the collection of NIAID PaVE (https://pave.niaid.nih.gov) [21] were downloaded as GenBank files. The files were imported into CLC and customized for use as databases. Customization involved manual entry of author-defined, clinically relevant, common 7-level taxonomic nomenclature into each HPV genome file. Due to the lack of consensus in virus taxonomy and recent changes by the ICTV, our nomenclature was based on the PaVE classification with attributes from the Baltimore classification and the original ICTV 5-rank and current 15-rank taxonomic structures [19]. Specifically, we defined our 7-level taxonomic ranks as: Virus_nucleic acid type; Family; Genus; Species; Type; Lineage; and Sublineage. For HPV reference genomes, the taxonomic nomenclature was annotated to the 5th or "Type" level. For HPV variant genomes, the taxonomic nomenclature was annotated to the 7th or "Sublineage" level. For example, the taxonomy of HPV-16 reference genome was "Virus_dsDNA; Papillomaviridae; Alpha; Alpha 9; HPV16; blank; blank." The taxonomy of an HPV-16 variant genome was defined to the sublineage level i.e., "Virus_dsDNA; Papillomaviridae; Alpha; Alpha 9; HPV16; A; A1." Collectively, the taxonomized genome sequences were converted to 3 distinct CLC databases: 1) HPV Taxonomic Profiling Index, 2) HPV Sequence List, and 3) HPV BLAST database (Figure 6A). The database creation tools employed for this function were: Create Taxonomic Profiling Index, Sequence List, and Create BLAST Database. Finally, the three HPV reference databases were integrated and utilized within its respective workflows or tools: 1) Taxonomic Profiling, 2) Map Reads to Reference, and 3) multi-BLAST tool (Figure 6A).



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2021 doi:10.20944/preprints202108.0149.v1

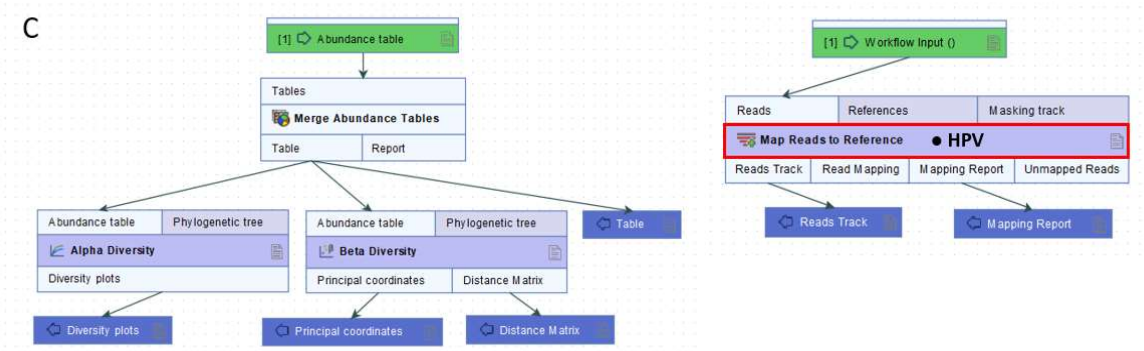


Figure 6. CLC Workflows, Tools, and Databases. (**A**) Microbial Genomics Module containing ready-to-use workflows and tools for abundance analysis (left). Customized HPV databases adapted from the PapillomaVirus Espiteme (PaVE) reference genomes (n = 219) for use in CLC (right). (**B**) Data QC and Taxonomic Profiling workflow incorporating the *HPV Reference Index* (● HPV) produces QC reports and HPV abundance tables from NGS reads of clinical samples. The Abundance Table output file ( ) is utilized as the input file for downstream diversity analysis shown in (C). (**C**) Merge and Estimate Alpha and Beta Diversities workflow generates diversity plots and statistical results (left). Map Reads to Reference workflow incorporating the *HPV Sequence List* (● HPV) generates an alignment map of the reads on the reference HPV genomes (right).

4.3. Data quality control (QC) and taxonomic profiling workflow

CLC Genomics Workbench 21.0.4 with CLC Microbial Genomics Module 21.0 (Redwood City, CA) were installed on an HP notebook computer (specifications: Intel i7-7500U dual-core processor @ 2.70 GHz and 8 GB RAM) for use throughout this project. The Microbial Genomics Module has various pre-built or "ready-to-use" workflows and tools as shown in Figure 6A. The pre-built "Data QC and Taxonomic Profiling" workflow with data input/output, individual elements (processing steps) and flow directions (arrows) are presented in Figure 6B. The analysis consisted of four primary steps: Data import, Data QC, Taxonomic Profiling, and Visualization of Abundance Table. Each step is described in detail here. First, fastq files generated after sequencing were imported as paired-end (forward-reverse) and merged using the "Import Illumina" tool for placement into a new "input" folder created within the file structure. Second, the workflow was initiated by clicking "Run" to carry out these steps: 1) select input files, 2) preprocess reads with quality trimming based on quality scores with a limit cutoff 0.05, and the ambiguity number ≤2, and adapter trimming, 3) map and assign reads to the HPV reference index, and 4) quantify the abundance of each qualified HPV genotype to generate an abundance table for each sample.

The post-workflow finishing steps are described here. First, the resultant taxon tables were merged into one using the "Merge Abundance Tables" tool, cleaned by filtering out unmapped (genotype "Unknown") sequences, and saved as a clean subtable. Second, the merged (filtered) taxon table was joined by clinical metadata (.xlsx format) with the "Add Metadata to Abundance Table" tool. Third, the two (merged, filtered, metadata-added) tables from LSIL (n = 95) and HSIL (n = 60) groups of samples were merged into one (n = 155) named "LSIL_HSIL abundance table" and aggregated by feature i.e., "Family" using the "Aggregate feature" option of the table (Supplementary Table 1). This option was applied to avoid lengthy feature names in the abundance table and graphs.

4.4. Estimate alpha and beta diversities workflow

The pre-built "Merge and Estimate Alpha and Beta Diversities" workflow (Figure 6C) was used to analyze and compare HPV communities between groups, such as, cytological categories. The analysis consisted of three primary steps: Abundance table import and merge, Alpha Diversity and Beta Diversity analyses with visualization of diversity

 doi:10.20944/preprints202108.0149.v1

plots. The workflow was initiated by clicking "Run" to carry out these steps: 1) select cleaned LSIL and HSIL abundance tables with appended metadata, 2) merge abundance tables, 3) compute α-diversity of the HPV communities to measure within-sample variation, and 4) compute β-diversity of the HPV communities to measure between-sample variation. The choices for α-diversity measures included: Total number, Chao 1 bias-corrected, Chao 1, Simpson's index, Shannon entropy, and Phylogenetic diversity. For this study, two α-diversity measures were computed, i.e. Simpson's index [28]: SI = $1 - \sum_{i=1}^{n} p_i^2$, and Shannon entropy [29]: $H=\sum_{1}^{n} p_i \log_2 p_i$, where n was the number of HPV genotypes found in the sample, and p_i was the proportion of reads that were identified as the i^{th} HPV genotype. For β-diversity measures, the choices included: Bray-Curtis, Jaccard, Euclidean, and Phylogenetic diversity (UniFrac method with 5 variations). For this study, β-diversity was measured by Bray-Curtis distances or compositional "dissimilarity" between samples [30]: $\mathcal{B} = \frac{\sum_{i=1}^{n} |x_i^A - x_i^B|}{\sum_{i=1}^{n} (x_i^A - x_i^B)}$, where n is the number of operational taxonomic unit (OTU) i and x_i^A and x_i^B are the respective abundances of OTU i in samples A and B, to measure the dissimilarity of HPV genotype composition between samples. Principal coordinate analysis (PCoA) was performed to determine and 3D-plot the correlative relationship between variables (HPV genotypes) in the LSIL or HSIL groups. The permutational multivariate analysis of variance (PERMANOVA) tool was used to test for statistical differences in the centroids and dispersion of the groups.

4.5. Differential abundance analysis methods

For differential abundance analysis of HPV communities between LSIL and HSIL groups, the "Convert Abundance Table to Experiment" and "Proportion-based Statistical Analysis" tools were used in succession. First, the merged LSIL and HSIL abundance table with appended metadata was chosen as input for conversion. The metadata group named "PAP" with two categorical variables "LSIL" and "HSIL" was selected as the factor. The output produced a table labeled "experiment" which was entered into the "Proportion-based Statistical Analysis" tool as input. The weighted proportion test of Baggerly [31] was chosen for comparing proportions of a two-group experiment. Fundamentally, the test compares counts, such as, read counts in relation to total sum of counts in each sample. By comparing weighted proportions instead of counts, the data is corrected for sample size. For this study, changes (up- or downtrends) in HPV communities with disease progression from LSIL to HSIL was our primary interest. For visualization of differences in HPV communities, the "Create Heat Map for Abundance Table" tool was used to generate a heat map from the merged LSIL and HSIL abundance table. Heat map parameters were set with 1-Pearson correlation as the distance and Average linkage as the clustering method.

4.6. Map reads to reference workflow

The "Map Reads to Reference" workflow (Figure 6C) maps each sequencing read against the reference HPV Sequence List. The mapping consisted of three primary steps: import paired-end reads, map reads to reference genomes, and create reads track for visualization. The workflow was initiated by clicking "Run" to carry out these steps: 1) select paired-end reads, 2) map reads to reference, and 3) create reads track as output.

5. Conclusions

CLC workflows with integrated, customized HPV reference genomes proved to be an ultra-fast and accurate method of HPV virome genotyping and profiling. By forging a path through the bioinformatics pipeline of a user-friendly software, the impasses to HPV discovery and beyond have been overcome.

Supplementary Materials: Figure S1: Correlation between BLAST statistics, Table S1: LSIL and HSIL merged abundance table, Table S2: HPV E6/E7 Sanger and deep sequencing results, Table S3: BLAST results for HPV genotyping by Sanger sequencing.

Author Contributions: Conceptualization, J.SG. and Y.W.; methodology, J.SG. and Y.W.; database, J.SG.; validation, J.SG. and Y.W.; formal analysis, J.SG. and Y.W.; investigation, J.SG., Q.X.; resources, J.SG. and Y.W.; data curation, J.SG., Q.X., H.C., and Y.W.; writing – original draft preparation, J.SG. and Y.W.; writing – review & editing, J.SG., Q.X., H.C. and Y.W.; visualization, J.SG. and Y.W.; supervision, J.SG. and Y.W.; project administration, J.SG. and Y.W.; funding acquisition, J.SG. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the DoD Congressionally Directed Medical Research Programs (CDMRP) grant to J.SG (#CDMRPL-16-0-DM160469), and the APC was funded by the Department of Clinical Investigation Intramural Funding Program at Brooke Army Medical Center, Fort Sam Houston, Texas.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the institutional review board of Brooke Army Medical Center (protocol code C.2017.090d, approval date June 9, 2017).

Informed Consent Statement: Patient consent was waived due to research involving no more than minimal risk to subjects, the waiver will not adversely affect the rights and welfare of the subjects, and research could not practicably be carried out without the waiver.

Data Availability Statement: The data presented in this study are openly available in the NCBI Sequence Read Archive (SRA). Title: HPV viromes in pap smears. SRA Accession Number: SRP323861; BioProject: PRJNA737277; BioSample Accession Numbers: SAMN19686938 to 19687092.

Acknowledgments: We thank the staff, Ms. Roxanne Toscano and Ms. Rosalyn Miller, at the Cytopathology Laboratory of Brooke Army Medical Center for their invaluable service for collecting the clinical samples in support of the HPV Research Program in the Dept. of Clinical Investigation at Brooke Army Medical Center. We also thank Brandon Converse for his technical and sequencing support at Lucigen Corporation, Middleton, WI.

Conflicts of Interest: The Defense Health Agency (DHA) of the U.S. Department of Defense has licensed the customized HPV database described herein to QIAGEN Digital Insights. The inventor of the customized taxonomy is J.SG. No potential conflicts of interest were disclosed by the other authors. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Author's Note: This paper has undergone PAO review at Brooke Army Medical Center and was cleared for publication. The view(s) expressed herein are those of the authors and do not reflect the official policy or position of Brooke Army Medical Center, the United States Army Medical Department, the United States Army Office of the Surgeon General, the Department of the Army, the Department of the Air Force, or the Department of Defense or the United States Government.

References

1. Mammas IN, Spandidos DA. Four historic legends in human papillomaviruses research. J BUON. 2015 Mar-Apr;20(2):658-61.
2. Dürst M, Gissmann L, Ikenberg H, zur Hausen H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc Natl Acad Sci U S A. 1983 Jun;80(12):3812-5. doi: 10.1073/pnas.80.12.3812.
3. Zur Hausen H. Cancers in Humans: A Lifelong Search for Contributions of Infectious Agents, Autobiographic Notes. Annu Rev Virol. 2019 Sep 29;6(1):1-28. doi: 10.1146/annurev-virology-092818-015907.
4. Javier RT, Butel JS. The history of tumor virology. Cancer Res. 2008 Oct 1;68(19):7693-706. doi: 10.1158/0008-5472.CAN-08-3301.
5. International Agency for Research on Cancer [IARC] (2012). Monographs on the Evaluation of Carcinogenic Risks to Humans-Human Papillomaviruses. Geneva: World Health Organization, 255–313.
6. Mastoraki A, Schizas D, Gkiala A, Ntella V, Hasemaki N, Pentara I, Ntomi V, Kapelouzou A, Liakakos T. Human Papilloma Virus infection and breast cancer development: Challenging theories and controversies with regard to their potential association. J BUON. 2020 May-Jun;25(3):1295-1301.
7. Liyanage SS, Rahman B, Ridda I, Newall AT, Tabrizi SN, Garland SM, Segelov E, Seale H, Crowe PJ, Moa A, Macintyre CR. The aetiological role of human papillomavirus in oesophageal squamous cell carcinoma: a meta-analysis. PLoS One. 2013 Jul 24;8(7):e69238. doi: 10.1371/journal.pone.0069238.
8. de Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. Int J Cancer. 2017 Aug 15;141(4):664-670. doi: 10.1002/ijc.30716.

9. Wild C. P., Weiderpass E., Stewart B. W. (2020). World Cancer Report: Cancer Research for Cancer Prevention. Lyon: International Agency for Research on Cancer.
10. Willemsen A, Bravo IG. Origin and evolution of papillomavirus (onco)genes and genomes. Philos Trans R Soc Lond B Biol Sci. 2019 May 27;374(1773):20180303. doi: 10.1098/rstb.2018.0303.
11. Bravo IG, Alonso A. Mucosal human papillomaviruses encode four different E5 proteins whose chemistry and phylogeny correlate with malignant or benign growth. J Virol. 2004 Dec;78(24):13613-26. doi: 10.1128/JVI.78.24.13613-13626.2004.
12. Chen Z, DeSalle R, Schiffman M, Herrero R, Wood CE, Ruiz JC, Clifford GM, Chan PKS, Burk RD. Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. PLoS Pathog. 2018 Nov 1;14(11):e1007352. doi: 10.1371/journal.ppat.1007352.
13. Dube Mandishora RS, Gjøtterud KS, Lagström S, Stray-Pedersen B, Duri K, Chin'ombe N, Nygård M, Christiansen IK, Ambur OH, Chirenje MZ, Rounge TB. Intra-host sequence variability in human papillomavirus. Papillomavirus Res. 2018 Jun;5:180-191. doi: 10.1016/j.pvr.2018.04.006.
14. Shen-Gunther J, Wang Y, Lai Z, Poage GM, Perez L, Huang TH. Deep sequencing of HPV E6/E7 genes reveals loss of genotypic diversity and gain of clonal dominance in high-grade intraepithelial lesions of the cervix. BMC Genomics. 2017 Mar 14;18(1):231. doi: 10.1186/s12864-017-3612-y.
15. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1-8. doi: 10.1016/j.ygeno.2015.11.003.
16. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, Morton L, Jarman RG. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. J Infect Dis. 2020 Mar 28;221(Suppl 3):S292-S307. doi: 10.1093/infdis/jiz286.
17. Ladoukakis E, Kolisis FN, Chatziioannou AA. Integrative workflows for metagenomic analysis. Front Cell Dev Biol. 2014 Nov 19;2:70. doi: 10.3389/fcell.2014.00070.
18. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated Omics: Tools, Advances, and Future Approaches. J Mol Endocrinol. 2018 Jul 13:JME-18-0055. doi: 10.1530/JME-18-0055.
19. International Committee on Taxonomy of Viruses Executive Committee. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol. 2020 May;5(5):668-674. doi: 10.1038/s41564-020-0709-x.
20. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018 May 17;6(1):90. doi: 10.1186/s40168-018-0470-z.
21. Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, Sun Q, Kaur R, Huyen Y, McBride AA. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. 2017 Jan 4;45(D1):D499-D506. doi: 10.1093/nar/gkw879.
22. From humble tool to global icon. news.bbc.co.uk. Available online: http://news.bbc.co.uk/2/hi/europe/8172917.stm. (Accessed 5 June 2021)
23. Latsuzbaia A, Wienecke-Baldacchino A, Tapp J, Arbyn M, Karabegović I, Chen Z, Fischer M, Mühlschlegel F, Weyers S, Pesch P, Mossong J. Characterization and Diversity of 243 Complete Human Papillomavirus Genomes in Cervical Swabs Using Next Generation Sequencing. Viruses. 2020 Dec 14;12(12):1437. doi: 10.3390/v12121437.
24. List of sequence alignment software. Wikipedia. Available online: https://en.wikipedia.org/wiki/List_of_sequence_alignment_software (Accessed 5 June 2021)
25. Arteche-López A, Ávila-Fernández A, Romero R, Riveiro-Álvarez R, López-Martínez MA, Giménez-Pardo A, Vélez-Monsalve C, Gallego-Merlo J, García-Vara I, Almoguera B, Bustamante-Aragonés A, Blanco-Kelly F, Tahsin-Swafiri S, Rodríguez-Pinilla E, Minguez P, Lorda I, Trujillo-Tiebas MJ, Ayuso C. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. Sci Rep. 2021 Mar 11;11(1):5697. doi: 10.1038/s41598-021-85182-w.
26. De Cario R, Kura A, Suraci S, Magi A, Volta A, Marcucci R, Gori AM, Pepe G, Giusti B, Sticchi E. Sanger Validation of High-Throughput Sequencing in Genetic Diagnosis: Still the Best Practice?. Front Genet. 2020 Dec 2;11:592588. doi: 10.3389/fgene.2020.592588.
27. Shen-Gunther J, Xia Q, Stacey W, Asusta HB. Molecular Pap Smear: Validation of HPV Genotype and Host Methylation Profiles of ADCY8, CDH8, and ZNF582 as a Predictor of Cervical Cytopathology. Front Microbiol. 2020 Oct 15;11:595902. doi: 10.3389/fmicb.2020.595902.
28. Simpson E. H. (1949). Measurement of diversity. Nature 163 688–688.
29. Shannon C. E. (1948). A mathematical theory of communication. Bell Syst. Techn. J. 27 623–656.
30. Bray J. R., Curtis J. T. (1957). An ordination of the upland forest communities of Southern Wisconsin. Ecol. Monogr. 27 326–349.
31. Baggerly KA, Deng L, Morris JS, Aldaz CM. Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics. 2003 Aug 12;19(12):1477-83. doi: 10.1093/bioinformatics/btg173.2. Dürst M, Gissmann L, Ikenberg H, zur Hausen H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions Proc Natl Acad Sci U S A. 1983 Jun;80(12):3812-5. doi: 10.1073/pnas.80.12.3812.
32. CLC Microbial Genomics Module User Manual: Taxonomic Profiling. QIAGEN Digital Insights. Available online: https://digitalinsights.qiagen.com/products-overview/plugins/(Accessed 6 June 2021)

 doi:10.20944/preprints202108.0149.v1

33. How to BLAST Guide. National Center for Biotechnology Information. Available online: https://ftp.ncbi.nlm.nih.gov/pub/fact-sheets/HowTo_BLASTGuide.pdf (Accessed 6 June 2021)