Master Degree Project

UNIVERSITY OF SKÖVDE

COMPUTATIONAL DETECTION OF HUMAN PAPILLOMA VIRUS IN THE CERVICAL CANCER GENOME

Master Degree Project in Bioinformatics

Second Cycle 30 credits
Spring term 2023

Hope Omoghene
a22hopom@student.his.se

Supervisor:
Zelmina Lubovac
zelmina.lubovac.@his.se

Examiner:
Sanja Jurcevic
Sanja.jurcevic@his.se

ABSTRACT

Global research and development have witnessed new horizons in technological advancements, especially in the use of new-generation bioinformatic tools to solve human needs. Cervical cancer, caused by a sexually transmitted virus like human papillomavirus (HPV), is one of the most common cancers threatening women's health. The main aim of the study is to evaluate existing Next-generation pipelines for detection of HPV in cervical cancer. The method includes data retrieval, which involves careful selection and downloading of 30 metagenomic data (in FASTA-Q format) from the Human Microbiome Project database. The implementation phase of the study involved setting up and configuring the virus detection tools (HPViewer, VirusSeq and VirusFinder 2.0). All the tools were run on default settings to analyze the metagenome samples using the instructions provided by their authors. The result showed that the tools detected HPV. The HPViewer demonstrated a higher level of HPV detection, followed by VirusSeq and then VirusFinder 2. The HPViewer had the shortest run time, completing an analysis in 24.1 seconds, followed by VirusFinder 2 in 208 seconds and VirusSeq took 4200 seconds (1 hour, 10 minutes to run). HPViewer demonstrated an outstanding sensitivity of 100%, VirusFinder 2 (45.5 %) and VirusSeq (63.6%). In conclusion, the present study underscored the trade-offs between speed, accuracy, and resource consumption between bioinformatics tools for HPV detection. Each of the tools exhibited unique strengths and limitations; however, they provided valuable options for HPV detection.

Keywords: VirusFinder2, HPViewer, VirusSeq, HPV, Bioinformatics, Cervical cancer

LIST OF ABBREVIATIONS

CNV - Copy Number Variation

DNA – Deoxyribonucleic Acid

HIVID – High-throughput Viral Integration Detection

HMP – Human Microbiome Project

HNSCC - Head and Neck Squamous Cell Carcinoma

HPV – Human Papillomavirus

NGS – Next-Generation Sequencing

TABLE OF CONTENTS

Contents

LIST OF TABLES

LIST OF FIGURES

Introduction

For the past decades, hitherto, global research and development has witnessed new horizons in technological advancements, especially in the use of new generation bioinformatic tools to solve human needs with respect to disease diagnosis and management. In the aspect of cervical cancer, there have been approaches targeted at early and accurate detection of the causative organisms (Human papillomavirus). The next-generation sequencing (NGS) has shown significant impacts in actualizing the profiling of HPV virus (Shen-Gunther et al., 2021). Over the years, wet lab experiments of genomics studies have been streamlined and made simpler via high throughput technology, but bioinformatics analysis, on the other hand, still has some element of bottleneck, hence the need to evaluate and re-evaluate the bioinformatics tools to validate their relevance in the 21st-century research and development. The use of NGS tools in HPV detection in cervical cancer is very critical and invaluable, especially in handling the complex and enormous data generated from the genome database. This approach is achievable because of the open-source tools which are available online, are usually command-line based and require coding skills (Shen-Gunther et al., 2021).

Problem Definition

The present study deals with cervical cancer and the detection of the causative organism (HPV) using bioinformatic tools. HPV detection involves a computational-based technique where NGS data are used to align the data against HPV multi-reference genome sequences. Cervical cancer is one of the most common cancers threatening women's health and is the fourth most common cancer in women worldwide (Pimple and Mishra, 2022). Currently, cervical cancer is caused by a sexually transmitted infection like human papillomavirus (HPV) (Hu *et al.*, 2015). HPV is a double-stranded DNA virus and its infection is the most prevalent sexually transmitted disease, resulting in over 14000000 individuals every year and 80% of sexually active individuals in their lifetime being infected by HPV (Hathaway, 2012; Sendagorta-Cudós and Burgos-Cibrián, 2019). HPVs, being double-stranded DNA, cannot be cultured; hence, their detection relies on a variety of techniques, such as molecular technique through sequencing, serology, and immunology, either in vitro or in vivo. New-generation sequencing (PCR-based) is the most current technique for HPV detection (Goswami, 2016; Hu and Ma, 2018; Arroyo Mühr *et al.*, 2020).

The recent increase in the availability of high-throughput data offers opportunities to study viral genetic material in the host genome. The next-generation sequencing (NGS) provides the privilege of

detecting viral species like HPV from GenBank data on human tissues (Li et al., 2013). But, since existing computational techniques do not optimally investigate clinical samples, according to Khan et al., (2019), PCR-based assays or bioinformatic methods can be adapted for the detection of the HPV using data from GenBank (Moreau *et al.*, 2013; Fraser *et al.*, 2019; Yan *et al.*, 2019). Thus, NGS is an informative tool that can guide health oncologists during cancer management and help in personalizing the treatment in cervical cancer (Bettoni *et al.*, 2017). By accessing GenBank and repositories, it is possible to use NGS novel tools to refine HPV diagnosis in cervical cancer and to predict the cancer response to specific anticancer drugs (Bettoni *et al.*, 2017). In this study, bioinformatic tools that have been designed for the evaluation of HPV expression in cervical cancer will be assessed and the comparison of their functionality will be done. NGS encompasses the use of the High-throughput Viral Integration Detection (HIVID) method (W. Li *et al.*, 2013; Augustin *et al.*, 2020). These novel bioinformatic techniques can detect the expression of HPV in the cervical cancer genome; it can also determine co-infection among the HPV types probed along with their integration sites (Li et al., 2013).

Problem Motivation

Despite numerous studies carried out on cervical cancer in recent times, the computational characterization of the etiological agent of cervical cancer (HPV) has not been fully elucidated and authenticated. Hence, the present study tries to find easier, faster, and cost-effective tools for detecting HPV. Prior to the applications of bioinformatic tools in oncology research, detections of HPV were basically carried out using nested PCR-based primers such as CPI/II and MY09/11 systems (Chandrani *et al.*, 2015a). Other approaches in the screening of cervical cancer viruses involve the use of methods such as signal-amplification assays, hybridization-based SPF and nucleic-acid-based approaches such as PCR-based, microarray and real-time techniques (Gates *et al.*, 2021). However, all these technologies used for detecting HPV have some limitations, such as cost, time of processing and specificity of result, contrary to NGS, where a single run can generate millions of data reads in 24 hours, which is far greater than what conventional methods can achieve (Nilyanimit *et al.*, 2018). Thus, conventional methods used before now have an inability to handle complex and enormous data needed to individualize the outcome within a short possible period. This formed the motivation and basis for adopting NGS techniques to solve oncological research needs.

Overview of Previous Research

There are a number of NGS research which adopted different tools for the detection of HPV in cervical cancer. Chandrani et al., (2015), in their study, designed HPV-detector, a bioinformatic tool with a graphic user interface (GUI)-based. It is solely for detection and annotation of cervical cancer HPV genome and the principle is based on NGS data sets. Chandrani et al., (2015), in their study, developed "a custom-made reference genome" made up of human chromosomes together with HPV annotated genome. The HPV-detector runs on a dual mode: a 'quick mode' and an 'integration mode'. The HPV-Detector developed by Chandrani et al., (2015) was made accessible in public domain for download using the link http://www.actrec.gov.in/pi-webpages/AmitDutt/HPVdetector/HPVDetector.html. The outcome of the study by Chandrani et al., (2015) showed that their NGS-tool was able to identify the presence of HPV in cervical cancer samples. The study thus concluded that HPV-Detector is a simple and precise tool that is very robust in detecting HPV from cervical cancer samples using whole genome, transcriptome and whole exome.

A similar study by Khan et al., (2019) was aimed at detecting viruses in neoplastic human tissues using RNA-seq data. They developed a bioinformatic method called VirTect for the virus detection. The VirTect was used to analyze RNA-seq data from 363 Head and Neck Squamous Cell Carcinoma (HNSCC) patients. From the outcome of the study, VirTect showed a better performance in accuracy compared to other existing prediction methods such as VirusSeq and VirusFinder. The study opined that carcinogenesis of HPV-induced HNSCC involves different genes. The outcome of their study showed that the NGS tool can be used as a prediction tool in cervical cancer diagnosis. This was validated manually through pathological findings on histopathologic specimens. Khan et al., (2019) posited that the VirTect tool showed the best performance in accuracy and recall when compared to other existing NGS prediction tools (VirusSeq and VirusFinder) with respect to identifying viral sequences from gene repository data. The study thus concluded that VirTect is an effective tool for the detection of viruses from cancer samples and can facilitate the characterization of various types of cancer.

Yan et al., (2019), in their study, developed a bioinformatic tool, DisV-HPV16, to investigate both HPV16 detection and gene expression. The NGS tool designed by Yan et al., (2019) differs from almost all of the existing bioinformatic tools, which are mostly used for data detection in viral infection or in genome integration. From the outcome of their findings, they were able to rapidly detect the HPV16 virus and viral oncogene expression using the DisV-HPV16. The study thus concluded that DisV-HPV16 tool was very convenient for HPV detection and that the accuracy of DisV-HPV16 was affirmed in laboratory experiments; hence, the tool is recommended for future research. The DisV-HPV16 tool was shown to be highly effective in virus detection after modification of the reference file. The accuracy of

this tool was affirmed empirically in the web lab. experiments. DisV-HPV16, according to their research findings, showed significant reliability of the protocols compared to other software. The study thus concluded that DisV-HPV16 is a novel tool for HPV detection and also detects viral oncogene expression through analysis of RNA sequencing data.

In a similar study by Qiu et al., (2022), which evaluated the genetic landscape of cervical cancer using a multigene next generation sequencing (NGS) panel, the study analyzed 64 samples of Chinese cervical cancer patients. The result identified about 810 somatic variants, 701 copy number variations (CNVs), and 2730 germline mutations. The NGS analysis revealed several genetic mutations in patients with cervical cancer and it also detected the *PIK3CA* gene in cervical cancer. To authenticate the functionality of the validated multigene NGS panel, the role of *PIK3CA* predicted in the cervical cancer cells was further compared with the ONCOKB database. From the analysis of their findings, it can be inferred that cervical cancer patients could benefit from PARP inhibitors. Generally, this study showed that genetic mutations affect the genetic susceptibility to cervical cancer.

In another similar research by Lee et al., (2020), the challenge of detecting HPV for patients that frequently undergo biopsies was addressed by developing a next-generation sequencing method that measures circulating HPV-DNA using panHPV-detect. The outcome of the study showed that in pre-CRT samples, panHPV-detect showed 100% specificity and sensitivity for HPV. It can thus be concluded from the study that PanHPV-detect demonstrated a very high outcome with respect to specificity, accuracy, and sensitivity in the identification of cHPV-DNA diagnosis.

Aims and Objectives

The main aim of the study is to evaluate existing next-generation pipelines for the detection of HPV in cervical cancer.

The objectives are:

To investigate the potential of different tools to be combined into a pipeline for HPV detection in cervical cancer.

To assess the level of accuracy of bioinformatic tools in detecting the presence of HPV

To compare the level of sensitivity and specificity of the selected tools

Relevant methods

There are a variety of bioinformatic tools designed for gene detection and integration in the human genome. Some of these tools include Virus Seq, Viral Fusion Seq, Virus Finder, SeqMap, ReadSCAN and RINS, among others (Chandrani et al., 2015; He et al., 2021; Naeem et al., 2013; Wang et al., 2013). All these bioinformatic tools have their different functions in line with what researchers need. They can detect the HPV sequence along with other viruses. SeqMap 2.0 is a web-based system that has been used in the past decades. It works by employing pre-defined viral features in order to locate the integration sites of the virus (Li et al., 2013). The framework of SeqMap is based on the 454-sequencing database. It is a very reliable tool, although it does not evaluate the putative fusion breakpoints, hence, the framework could not discover novel HPV integrations (Li et al., 2013). Based on this limitation, VirusSeq was introduced for virus species detection in sequence data with viral integration events using Read Pair (RP) information (Chen *et al.*, 2013; Visser, Burger and Maree, 2016). VirusSeq works via computational subtraction of human sequences by alignment of the raw pair-end reads from whole-genome or transcriptome sequence. Thus, it can effectively generate a set of non-human sequences by subtracting the human sequences. In the second step, VirusSeq bioinformatic tool works by aligning the non-human sequences against a gene bank that contains all known viral sequences from “Genome Information Broker for Viruses”. In this approach, any virus with an overall mapping below the cut-off (1000 reads) is treated as non-existent within a virus genome; this cut-off applies to both whole-genome and RNA-Seq data (Chen et al., 2013). DisV-HPV16 is a novel bioinformatic tool designed to detect HPV and viral gene expression via the analysis of sequence data from a public database. DisV-HPV16 software can be downloaded for use, and it can yield outcomes that are accurate with more comprehensive insights into the status of the cervical virus infection and the host cell genome. DisV-HPV16 software is very sensitive, fast in operation and accurate.

Materials and methods

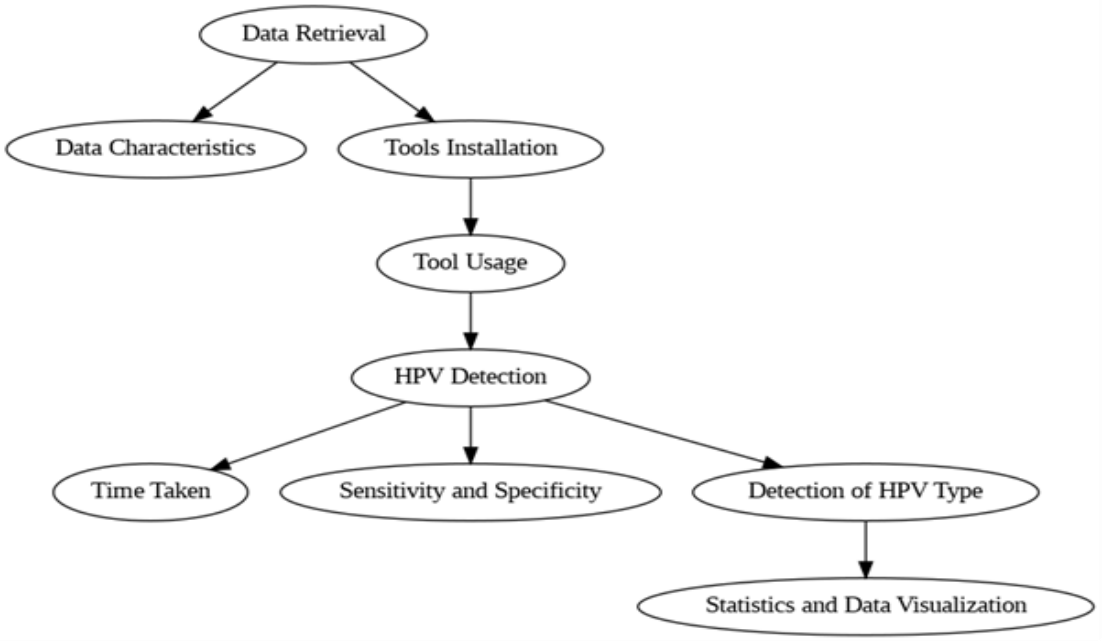

Figure 1: Flowchart of the methodology

The first step in the implementation of this study was the careful selection and acquisition of metagenomic data from reliable sources. Specifically, 30 metagenomic data samples were retrieved from the Human Microbiome Project (HMP) (https://hmpdacc.org/hmp/) (Table 1, Appendix). These data were selected based on their previous use in Hao et al. (2018), ensuring a reference point for comparing the HPV detection results obtained from the tools used in this study.

Understanding the metagenomic data used in this analysis is crucial, as it forms the foundation for assessing the efficacy of HPV detection tools. Key characteristics of the data that would be extracted include:

Data Format: The format of the data is fasta, fastq or fastq.gz.

Data Size: The size of the metagenome, both in byte and base pair.

Body Site: the body site from which the data was extracted.

Choice of Method

Conventionally, the detection of HPV relied on two assays: the rapid high-throughput target amplification (RHTA) and the broad-spectrum signal amplification (BSA) assay. However, in the current digital era, next-generation sequencing (NGS) has brought in a paradigm shift, going beyond simple

detection of the presence or absence of HPVs to provide thorough insights on the profiling of HPV sequences and specific infections. This research is concerned with the computational detection of HPV using different molecular biological information related to cervical cancer (Helene and Francois, 2013). Most of the bioinformatics studies involve analyzing biological data or information following the large and complex data generated in research. Most of the recent bioinformatics research deals with functional and structural aspects of proteins and genes (Pongor and Landsman, 2015). All the areas of modern biological and medical sciences make use of computational techniques; thus, molecular biology (genetic engineering) can't exist without bioinformatic tools (Markowetz, 2017; Lanigan, Kopera and Saunders, 2020). Developed countries have easy access to bioinformatics due to the availability of computer infrastructure and software knowledge. However, the use of this technology is premised on bioinformatics knowledge available in the public domain (Mulder *et al.*, 2018). After performing a comprehensive literature search, there are a limited number of studies on the computational detection of HPV in the cervical cancer genome using bioinformatic tools. Only a few of them have been designed and used in HPV detection; hence, the need to perform comparative studies in this field that will compare the effectiveness of some selected tools such as HPV-Detector, VirusSeq and VirusFinder-2 in detecting HPV cervical cancer genome data (Wang, Jia and Zhao, 2013; Chandrani *et al.*, 2015b; Khan *et al.*, 2019).

The tools that were initially chosen for evaluation were HPV detector, HPV meta and virus finder 2. However, HPV detector required permission from the developers before it could be downloaded, but only an auto generated response was given when the required form was filled, with no access to the tool. HPV meta authors included the github link to their tool in the publication, but there was no documentation on how to install or use the tool in the publication or on the github page. Efforts were made to find solutions to these challenges on third-party sites (e.g., bioinformatic forum sites), but no satisfactory solution was found, hence the need for the selected tools.

Method Description

HPViewer

This is a bioinformatic tool developed by Hao et al., (2018). It is a novel HPV detector. HPViewer is a specialized tool designed for the detection of HPV genome in metagenomic data. It can detect the presence and type. The installation process was relatively straightforward and well-documented, with clear instructions provided by the tool's developers.

HPViewer has been described by the author as a tool that can minimize false detection of HPV in the cervical cancer sequence. It does this by masking all the simple repeats that are common among the human genome and the homologous sequences that are common to different types of HPV. The study by Hao et al., (2018) ascertained the specificity and sensitivity of HPViewer tool using simulation samples. The result of their finding showed that the performance of HPViewer in detecting HPV in cervical cancer samples is more specific than other tools like VirusTAP and HPVDetector. It was concluded that HPViewer tool has the capacity to define the prevalence and distribution of HPV. The tool also explored the co-occurrence patterns of HPV at different sites in the body.

VirusSeq

VirusSeq, an essential component of the analysis, presented some installation challenges. It required a specific set of dependencies and software prerequisites that, at times, posed compatibility issues with the computing environment. After overcoming these challenges, the successful installation of VirusSeq was achieved, seamlessly integrating various viral detection methods, such as de novo assembly and reference-based alignment. VirusSeq works via computational subtraction of human sequences by alignment of the raw pair-end reads from the whole-genome or transcriptome sequence. Thus, it can effectively generate a set of non-human sequences by subtracting the human sequences. In the second step, VirusSeq bioinformatic tool works by aligning the non-human sequences against a gene bank that contains all known viral sequences from “Genome Information Broker for Viruses.” In this approach, any virus with an overall mapping below the cut-off (1000 reads) is treated as non-existent within a virus genome; this cut-off applies to both whole-genome and RNA-Seq data (Chen *et al.*, 2013).

VirusFinder 2.0

This NGS technique makes use of software applications for detecting intra-host viruses such as HPV via next generation sequencing (NGS) data (Wang, Jia and Zhao, 2013). VirusFinder has some level of specificity; it can detect virus infection, virus integration sites, co-infection associated with the viruses, and mutations in the genomes of the virus (Wang, 2015). VirusFinder has versions 1 and 2, with version 2 being the most current and an advancement of the former. VirusFinder V.2 can work with human genome data and can also work with other organisms aside from humans that have reference genome data available. The advantage VirusFinder 2 has over the former version is that it utilizes both single-end and paired-end data, unlike the former version 1, which can only work with paired-end reads. VirusFinder 2 can deal with the following types of NGS datasets: “whole transcriptome sequencing

(RNA-Seq)", "whole genome sequencing (WGS)," "ultra-deep amplicon sequencing" and "whole exome sequencing (WES)". According to Wang (2015), VirusFinder 2 also showed efficiency in the implementation of their new algorithm design (virus integration site detection through Reference Sequence customization). The rationale behind the VERSE algorithm is to harness virus detection by designing "personalized reference genomes." The outcome of the research by Wang (2015) showed that the VERSE algorithm in VirusFinder 2 significantly improved the sensitivity of virus detection at the integration site. VirusFinder 2.0 was an improvement of the version 1.0 since the functionalities were upgraded and the accuracy of the new version was improved based on the feedback from users of the first version. The new version of VirusFinder thus incorporated all the concerns raised about VirusFinder 1, and also provided novel functions that help detect viruses and characterize intra-host viruses. VirusFinder Version 2 can be accessed from https://bioinfo.uth.edu/VirusFinder/ and it requires Java 1.6 and Perl 5 to run. VirusFinder.pl scripts are available for users without programming skills. The script prepares input data for every step of the pipeline. It also processes the outputs of the analysis after the functions terminate. VirusFinder 2.0 does not need sequences of viruses as an input prerequisite. Hence, it can function efficiently with NGS data where the virus type is specified, for example, HPV in the present study.

Method Evaluation and Comparative Analysis

The study was segmented based on the following data retrieval characteristics, tools Installation, tools usage, detection (the ability of the different tools to detect the HPV), the ability of the different tools to detect the specific type of the HPV, the time taken for the different tools to run per sample, their sensitivity level, and their specificity. The visualized image will further evaluate each of the tools.

The selected methods above have been reported by previous studies to have application in the present study that deals with HPV detections, hence the comparative analysis of the tools.

VirusSeq Applications

The genome sequences of viruses in VirusSeq are well-known in terms of cervical cancer association and were detected in the detection step. The Cancer Genome Atlas (TCGA) dataset was combined into a single chromosome called chrVirus, and each viral gene's associated annotation was formatted in refFlat. A new hybrid reference genome named hg19Virus was built by combining hg19 and chrVirus. All Paired-End (PE) reads without computational subtraction are mapped to this reference (hg19Virus). If the PE reads are uniquely mapped with one end to one human chromosome and the other to chr25, the read pair is reported as a discordant read pair. All discordant reads are then annotated with human

and viral genes defined in the curated refFlat file. VirusSeq then clusters the discordant read pairs that support the same integration (fusion) event (e.g., HBV-MLL4). VirusSeq implements a dynamic clustering procedure. To remove outliers within a cluster, VirusSeq implemented the robust 'extreme studentized deviate' multiple-outlier detection procedure (Chen et al., 2013). Once outliers are detected within a cluster, the cluster boundary is reset by excluding the outlier reads. Analyzing the data within each cluster comes next after cluster boundaries have been established. In order to gather information and make inferences, this analysis looks at the traits and trends of the data points inside the cluster. VirusSeq can detect known and novel virus-human fusion events associated with diseases such as cervical cancer.

Virus Finder 2 Applications

Either raw next-generation sequencing reads in Fastq format or an aligned sequence file in BAM format can be imported into VirusFinder 2. If you input only raw reads, VirusFinder uses the Bowtie 2 aligner to quickly align them to reference genomes from GenBank, which can be accessed via (http://www.ncbi.nlm.nih. or http://hgdownload.cse.ucsc.edu/downloads.html). After that, for comparative downstream analysis, VirusFinder 2 aligns all reads—whether they were initially supplied aligned or aligned during the pipeline—to the human reference genome. For instance, VirusFinder 2 skips its viral identification step—which would typically align unmapped reads to an extensive virus genome database for de novo detection—when an input contains reads of the cancer-associated human papillomavirus (HPV). This database alignment is omitted because it is known that the virus is HPV. Rather, HPV-mapped reads are assembled straight into sequence contigs by the process. Subsequently, these contigs are mapped independently to the human reference and viral genome databases, which can be accessed via http://gib-v.genes.nig.ac.jp/. Strong human alignment contigs are removed. To find the existence and quantity of the viral sequences, VirusFinder rates and ranks the remaining HPV contig alignments.

In contrast to RINS described by Wang et al., (2013), which identify non-human sequences by analyzing reads unmapped to both virus and human references, VirusFinder 2 concentrates the analysis on reads that are known to be virus-mapped, offering speed and accuracy while streamlining the process.

HPViewer Application

This bioinformatics tool is applied in the masking approach of HPV detection to minimize the effect of the shared gene sequences during genotyping techniques (Hao et al., 2018). Usually, the shared sequences are filtered by aligning large raw reads from a sample, but, with the aid of the HPViewer, the repeat sequences in the reference HPV genome database can be masked. The masked HPV genomes are compared with that of human to ascertain any matches, which in turn indicate the elimination of false positive results.

The sensitivity and specificity of the tools were calculated according to Wu et al., (2023).

The formulas used were

$$Sensitivity = \frac{True\ positives}{True\ positives\ +\ False\ Negatives} \times 100$$

$$Specificity = \frac{True\ negatives}{True\ negatives\ + False\ positives} \times 100$$

True positive of a tool is defined as the number of samples that are deemed positive by the tool that are actually positive, according to (Hao et al., 2018). True negatives are the number of samples that are deemed negative by the tool that are actually negative according to the reference, while false positives and false negatives are the number of samples that are deemed positive or negative, respectively, by the tool but have a different result according to the reference (Hao et al., 2018).

Alternative Methods

There are a variety of bioinformatic tools designed for gene detection and integration in the human genome. Some of these tools include DisV-HPV16, Virus Seq, Viral Fusion Seq, Virus Finder, SeqMap, ReadSCAN and RINS, among others (Chandrani et al., 2015; He et al., 2021; Naeem et al., 2013; Wang et al., 2013). All these bioinformatic tools have their different functions in line with researchers' needs, although they are not HPV-specific. They can detect the HPV sequence along with other viruses, but the gap in these techniques is that the techniques lack the ability to annotate the region of the HPV genome detected (Chandrani et al., 2015; He et al., 2021; Naeem et al., 2013; Wang et al., 2013). Also, some of the tools can only operate on Linux platform, some are not functional in an operating system like windows.

DisV-HPV16 bioinformatic tool was designed by Yan et al., (2019). The rationale for their choice over other bioinformatic tools was that DisV-HPV16 has additional features which make it highly reliable bioinformatic software based on their study outcome. DisV-HPV16 is a novel bioinformatic tool designed to detect HPV and viral gene expression via the analysis of sequence data from public databases. DisV-HPV16 software is very sensitive, fast in operation and accurate. The processing DisV-HPV16 inputs, which include pair-end reads or raw single-end are usually converted into Fastq format to be mapped to a human reference genome with the aid of the HISAT alignment tool. The results will be sorted by SAM tools, and they will be annotated by StringTie. This step will show if the sample is HPV16 positive. The output result usually contains FPKM oncogene values, which can be used to determine the expression levels (Yan et al., 2019). However, the installation of the tool was not possible, as the author did not provide the link to download and install the tool.

In the era of artificial intelligence (AI), where bioinformatic tools are applied in various facets of life, especially in disease diagnosis, validated and accurate pipelines are very essential and critical in detecting HPV in order to understand and profile human papillomavirus in relation to cervical cancers. In a very recent discovery by Ure et al., (2022), who designed an open-source pipeline, “HPV-meta”, for the detection of HPV transcripts from sequence data, The “HPV-meta” pipeline is an automated system which can perform multiple steps at the same time. The functions that can be performed with the tool include HPV detection, human genome filtering, quality trimming, and generating fasta sequences for HPV positive samples. Fasta sequences can then be aligned to assess sequence diversity among HPV positive samples. HPV-meta has been used in identifying different types of HPV present in specimens. The “HPV-meta” pipeline is an efficient and validated pipeline that detects HPV by obtaining the fasta sequence file format.

SeqMap 2.0 is a web-based system used in the past decades. It works by employing pre-defined viral features in order to locate the integration sites of the virus (Li et al., 2013). The framework of SeqMap is based on the 454-sequencing database. It is a very reliable tool based on its efficacy in detecting sequence data, although it does not evaluate the putative fusion breakpoints; hence, the framework could not discover novel HPV integrations (Li et al., 2013). Based on this limitation, VirusSeq was introduced for virus species detection in sequence data with viral integration events using Read Pair (RP) information (Chen et al., 2013; Visser, Burger and Maree, 2016).

HPV-Detector is a specific bioinformatic tool that can detect multiple HPV types (Chandrani et al., 2015c). It can annotate and determine HPV integration sites utilizing whole-genome data, transcriptome, or raw exome as input. The HPV-detector can thus detect the presence of HPV sequences together with other virus types. However, HPV-detector lacks information to annotate the

position of the virus genome detected. HPV-Detector is a bioinformatics tool that is a user-friendly and unique tool used for analyzing Next Generation Sequencing data to detect HPV sequences from the cervical cancer genome. According to Chandrani et al., (2015), the HPV-Detector tool has been tested and validated, as it was able to detect 55 integration points in the cervical exome and it also detected neck and head transcriptome data sequences. The HPV Detector has shown an ability to perform a thorough sequence analysis in order to unveil the necessary information for co-infections associated with HPV subtypes among cervical cancer patients. The outcome of the study by Chandrani et al., (2015) showed there was a significant enrichment of the viral gene reads across the cervical cancer samples. This was achieved using the annotation module that is in-built in the HPV Detector tool. The result showed consistency with the existing biology of HPV genes and their functional activities in cervical cancer. The in-built annotation in the HPV Detector tool is a unique feature module that can help to demystify the role of other HPV open reading frame (ORF). Despite the restriction of cervical cancer analysis being restricted to its exome data sequence, a whole gene spectrum of the viral load present in any sample can also be run using the whole-genome data input. In the method, GenBank files of HPV types from a “web resource Papillomavirus Episteme (PAVE)” were accessed. The GenBank (.gb) files were converted into Fasta files. These reference sequences were composed of a multi-fasta sequence using bio-perl modules. The HPV genes were generated by parsing the GenBank (.gb) files. For the detection of HPV types, the multi-fasta HPV reference file was done using BWA aligner, after which read alignment was also done to index the virus genome. The aligned reads were generated from http://broadinstitute.github.io/picard. The human reference sequencing data was downloaded from http://hgdownload.cse.ucsc.edu/downloads. The file containing the entire HPV genome sequence was downloaded from the National Centre for Biotechnological Information (NCBI) database. The sequence data was converted into Fastq format to initiate E6 at position 104 of the sequence. HISAT2 was used to build reference files for the human sequence and the converted HPV sequence. RNA sequencing data from 18 HPV-positive patients was downloaded from GenBank. The sequence essence of the HELA and SIHA cell line will be downloaded also from GenBank. To assess the sensitivity and specificity of the HPV Detector, SiHa whole-genome sequence was downloaded from “Sequence Read Archive database of DDBJ'' using the link: https://trace.ddbj.nig.ac.jp/DRASearch/. The dataset was converted from SRA format to FASTQ format using the SRA tool kit. The FASTQ format files prepared were used to test the HPV detection using HPV-Detector.

Tools Installation

The implementation phase of the study involved setting up and configuring the virus detection tools. Initially, three tools were selected for evaluation: HPV Detection, HPV meta, and VirusFinder 2.

However, challenges were encountered with these tools, leading to a change in tool selection. Detailed information regarding tool installation, availability of documentation, language, installation difficulty, and requirements/dependencies is provided in results.

HPViewer (Hao et al., 2018)

In the implementation phase, the capabilities of HPViewer, a specialized software tailored for the precise detection of the Human Papillomavirus (HPV) genome within metagenomic datasets, were explored. Experience with HPViewer revealed an intuitive and straightforward installation process, thoughtfully documented by the tool's developers.

One notable aspect experienced was the tool's minimal system requirements and dependency needs, rendering it highly accessible and easy to integrate into the workflow. HPViewer user-friendly approach extends to its operation, where it offers a streamlined experience with a single command for execution. This simplicity enhances the efficiency of the HPV detection process, making it a valuable asset in my implementation efforts.

VirusSeq (Chen et al., 2013)

VirusSeq, an essential component of the analysis, presented some installation challenges. It required a specific set of dependencies and software prerequisites that, at times, posed compatibility issues with the computing environment. While VirusSeq didn't have as much dependencies as VirusFinder 2 (its main third-party tools were the Mosaik Suite), this suite required a significantly large amount of computing resources to run, most especially when making the reference 'jump' database. Hence, Google Colab (Google Colaboratory, n.d.) and kaggle (Kaggle, n.d.) were employed to run the tool. Google Colab is a cloud-based platform provided by Google that allows users to write and execute Python code in a web browser, while Kaggle is an online platform and a cloud-based workbench for data analysis and machine learning tasks. The 'jump' files were the index equivalent of VirusFinder 2, The special database called "jump" crafted by MosaikJump is like a helpful guide for efficiently matching sequencing reads to a reference genome in the Mosaik Aligner suite. The hash size was also reduced to 6 when making the jump files so as to reduce the memory requirements. The Mosaik suite bundled with the tool was not installable and ran independently and this caused a few issues with system permissions and restrictions, but they were rectified using the portable app version. When all the requirements were satisfied, the analysis was run. Running VirusSeq consists of entering multiple commands that perform different processes in analyzing the isolate. It should be noted that in order to get the actual result of the analysis, the log files were used. This was because the cutoff point of the

algorithm that selects from the log file the viruses to which there was alignment was too high (1000 alignments). The system chooses only the top virus match, even when this threshold is lowered. Because of this, the final VirusSeq output for samples containing different HPV strains frequently only displays the dominant strain. One must manually review the log files in order to record every strain. All virus types that were matched in the pipeline—including those excluded from the final report because of low read counts—have alignment counts recorded in the log files. Additional HPV strains co-infecting a sample can be found using the alignment evidence in the logs by choosing a lower threshold, like 10 alignments.

VirusFinder 2 (Wang et al., 2013)

Installation of VirusFinder 2, another critical tool in the analysis, proved to be a bit more complicated than the others as a lot of dependencies had to be installed. This tool specializes in detecting viral sequences in metagenomic data. VirusFinder 2 had a heavy reliance on various dependencies and third-party tools. This heavy dependence on external interactions led to numerous bugs, including instances where input parameters in the VirusFinder 2 code were improperly configured. For example, there were issues with the input parameter passed to 'samtools sort.' The symbolic link method, initially used in the source code to reduce memory usage, didn't function correctly and had to be replaced with the conventional copy-and-paste method.

Additionally, the tools demanded significant memory resources, particularly when generating the reference index to which the input reads would be aligned. Due to this resource demand, this process was outsourced to Google Colab. After ensuring that all requirements and indexes were properly configured, the analysis was initiated for all test isolates, and this proceeded without significant issues.

To its credit, VirusFinder 2 did offer a straightforward run command that required minimal tweaking once initially set up. It also facilitated an efficient means of automating the analysis process for all the isolates.

It is noticed that the tools require certain similar dependencies, such as bowtie and samtools. Table 1 shows the characteristics of the three tools used in the analysis. It shows that VirusSeq and VirusFinder 2 have more in common than HPViewer.

System and Software Requirements

Each of the selected tools came with its unique system and software requirements. The setup of the computing environment was influenced by these requirements, ensuring compatibility with the

selected tools. While some tools were platform-agnostic, others were designed for specific operating systems. Thus, the setup of the computing environment was influenced by the compatibility constraints of these tools.

Tool Usage

All tools were run on default settings, as this reflects the most likely mode of use for users who may not be proficient in programming. A description of how each tool was used is provided, including any challenges or adjustments required during tool setup and operation.

HPV Detection

The tools were used to analyze the metagenome samples using the instructions provided by their authors on how to run them. The results were then compared with each other to determine the capabilities of the tools to detect HPV.

Run Time

The study recorded the time it took for each tool to complete the analysis for the samples, and the averages were calculated. This was done to compare the runtime and speed of the tools.

Detection of HPV Type

The study also analyzed the number of detections of each HPV type by each tool across all samples, providing insights into the tools' ability to detect specific HPV types. The aim of this is to determine and analyze the ability of a tool to detect each of the different HPV types. This was done by analyzing the number of times a HPV type was detected by a tool and comparing it with the other

Statistics and data visualization

The data obtained from the test was compiled using Microsoft Excel. The tool was also used in analysis of the data. The result was visualized using Microsoft Excel, Python programming language and R.

Implementation

Data

Data Retrieval: The first step was the careful selection and acquisition of metagenomic data from reliable sources. 30 metagenomic data from the Human Microbiome Project (https://hmpdacc.org/hmp/) were retrieved. These data were selected based on their use in Hao et

al., 2018 . This enables us to have a metagenomic dataset whose HPV result is already known and thus would serve as a reference point in comparison with the result obtained from the tools used in this study.

Understanding the metagenomic data that underpins this analysis is essential. The data provides the raw material for the investigation into the efficacy of HPV detection tools. Key characteristics of the data are summarized

Data Format: The metagenomic data was provided in the widely used FASTQ format. This format, derived from Illumina sequencers, is well-suited for high-throughput sequencing data and includes information about the sequencing quality and base calls.

Data Size: The size of the metagenomic data is of critical importance in determining the computational resources required for analysis. The dataset ranged from several gigabytes to terabytes, reflecting the vast diversity in sample sizes.

Body site: The dataset used were all from the anterior_nares region.

Details of the data retrieved can be found in Table 1.

Table 1: Characteristics of selected tools

Tool	Availability of documentation	Language	Requirements/ dependencies	Web Interface	Input data
HPV viewer	Yes	Python	Python (2.7+), Python packages (sys, getopt, subprocess), Bowtie2, SAMtools, SAMtools	No	Raw reads
Virus Seq	Yes	Perl	Mosaik suite, perl, spanner	no	Raw reads
Virus finder 2	Yes	Perl	Linux, Bash shell, Java 1.6, Perl 5, BLAT, iCORN, CREST, GATK, BLAST+, Trinity, SVDetect, SAMtools, BWA, Bowtie2	No	Raw reads

HPViewer is a command line tool that does not provide a Graphic User Interface. However, for someone with basic knowledge of Linux/bash, the tool won't pose much challenge as the instructions on how to use it are straightforward and require minimal tweaking of parameters except for inputting the read files and indicating the output name.

VirusFinder 2 runs on Perl, and the instructions on how to run the tool are found in the tool manual (https://bioinfo.uth.edu/VirusFinder/VirusFinder-manual.pdf?csrt=16073299185204319372).
Running the tool is relatively straightforward, though it requires a little bit of modification to the configuration file, mainly to reflect the peculiarities to the system environment the tool is being used in.
VirusSeq also runs on Perl. It is initiated using a set of commands that perform different functions in the analysis process. There was a need to adjust the parameters so as to manage the available computing resources.

Optimization of the Tools

VirusSeq: The virus database for VirusSeq was optimized to include only HPV sequences; all non-HPV sequences were removed. This edited database was then used to create the jump files using Mosaic Jump, to which the input sequences were aligned during analyses. This was done to enhance the specificity of the tool for HPV. It also ensures the reduction of the size of the jump file, which means analysis can run faster than if the non-HPV sequences were included.

The hashsize parameter during the creation of the jump file was also reduced from the default of 15 to 6. It reduces the computing resources the tool needs to create the file and the time it takes. It also reduces the size of the file, thereby enhancing speed during analysis. Reducing the hashsize of the jump files also enhances the sensitivity of the tool during analysis.
VirusFinder 2: For VirusFinder 2, the virus database was also edited to include only HPV sequences. This edited database was then used to create the virus index files. This ensures the specificity of the tool and also enhances the speed of analysis as the index file sizes are smaller.

Results and Analysis

HPV detection

In our analysis of the HPV detection capabilities of three bioinformatic tools, namely HPViewer, VirusFinder 2, and VirusSeq, in comparison to the reference result from Hao et al. (2018), several key findings and trends emerged.

HPViewer exhibited results that closely aligned with the reference results, with a remarkable similarity except for four instances of false positives. This suggests that HPViewer demonstrates a high level of agreement with the reference outcome and can be considered a reliable tool for HPV detection.

In contrast, VirusFinder 2 demonstrated a lower overall performance, as it was only able to detect 5 positive results out of 11 samples. However, it did not produce any false positives, indicating that when it detects HPV, it is generally accurate. This suggests that while VirusFinder 2 may have a lower sensitivity, it exhibits a higher specificity in terms of avoiding false-positive results.

VirusSeq had 4 false positives and 4 false negatives, but had a closer detection pattern with HPV-Viewer and the reference than with VirusFinder 2. It had 7 true positives and 15 true negatives.

In Figure 2, we present a Venn diagram to visually depict the concurrence of detection results obtained by three distinct tools—HPViewer, VirusFinder_2, and VirusSeq. The intersections reveal the overlap in viral detections between these tools, with numerical annotations specifying the count of detections for each paired method. "Hao_et_al_2018" is used as a reference/benchmark to which the other tools were compared. Instances with no values indicate the absence of shared viral detections, while instances with values show that the overlapping tools have similar detections to the magnitude of the value in the overlapping region. This Venn diagram functions as a crucial instrument for the comparative evaluation of the detection capabilities and accuracy of the three methodologies. Furthermore, it aids in uncovering potential false positives or negatives within the dataset, using "Hao et al., 2018" as a benchmark.

VirusSeq log files show viral read alignments with HPV strains detected by the HPViewer pipeline. However, the main output only reports strains exceeding a conservative threshold of 1000+ supporting reads. Lowering the threshold to 10 aligned reads can identify additional HPV strains validated by HPViewer. By adjusting the main output threshold to 10 reads, VirusSeq can directly achieve higher sensitivity.

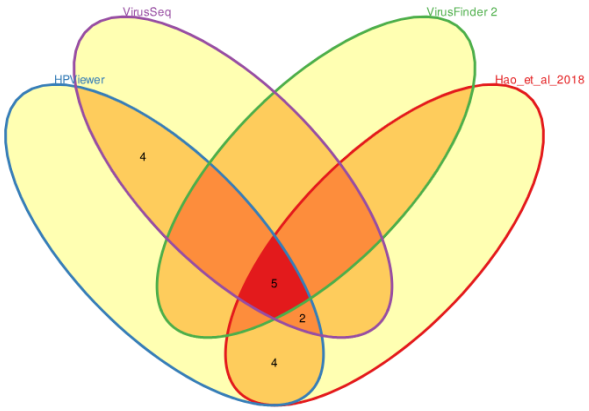

Figure 2: Venn Diagram showing the Intersection of the HPV Detection by the Tools. Instances with no values indicate the absence of shared viral detections, while instances with values show that the overlapping tools have similar detections.

Runtime

The time it took for the tools to complete the analysis of the samples was recorded, and the average taken (Figure 3). Only the time taken to analyze each sample was recorded. HPViewer had the shortest run time, completing an analysis in 24.1 seconds, followed by VirusFinder 2 in 208 seconds. VirusSeq takes 4200 seconds (1 hour, 10 minutes to run).

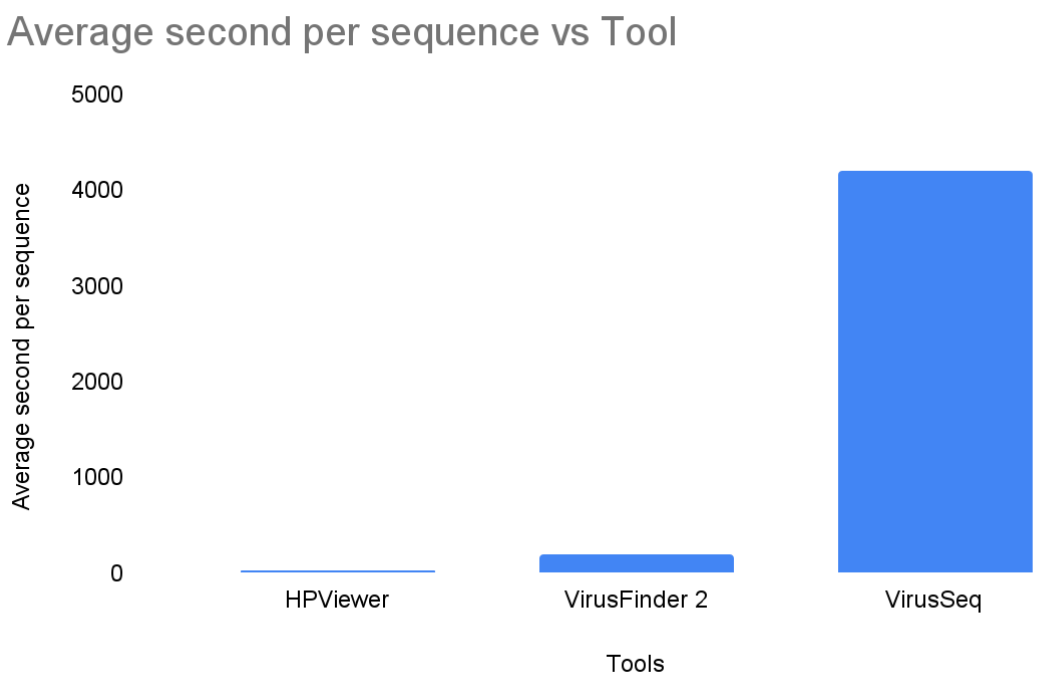

Figure 3: Time taken to run analysis on a sample per tool (Each bar represents the time in seconds taken by each tool).

As seen in figure 4, HPViewer demonstrated an outstanding sensitivity of 100%, signifying its exceptional ability to accurately identify all true positive cases, aligning with the reference data. This high sensitivity suggests that HPViewer effectively recognized all samples that genuinely contained HPV. However, its specificity was calculated at 78.9%, indicating a relatively higher rate of false positives. This implies that while HPViewer excelled in correctly identifying positive cases, it also had some instances of misidentifying negative cases.

Conversely, VirusFinder 2 exhibited a sensitivity of 45.5%, substantially lower than that of HPViewer. This suggests that VirusFinder 2 had difficulty identifying all true positive cases and likely missed several (Table 2). However, it demonstrated an impressive specificity of 100%, implying a negligible rate of false positives (Figure 4). VirusFinder 2 excelled in confirming the cases it identified as positive, but its lower sensitivity indicates a limitation in detecting all actual positive cases.

VirusSeq, the third tool, displayed intermediate values, with a sensitivity of 63.6% and a specificity of 78.9%. This indicates that VirusSeq successfully identified a significant portion of true positive cases but missed some (Table 2). Moreover, its specificity was relatively high, reflecting a lower rate of false positives compared to HPViewer.

Table 2: Detection accuracy of the three Tools

Tools	True positive	True Negative	False positive	False negative
HPViewer	11	15	4	0
VirusFinder 2	5	19	0	6
VirusSeq	7	15	4	4

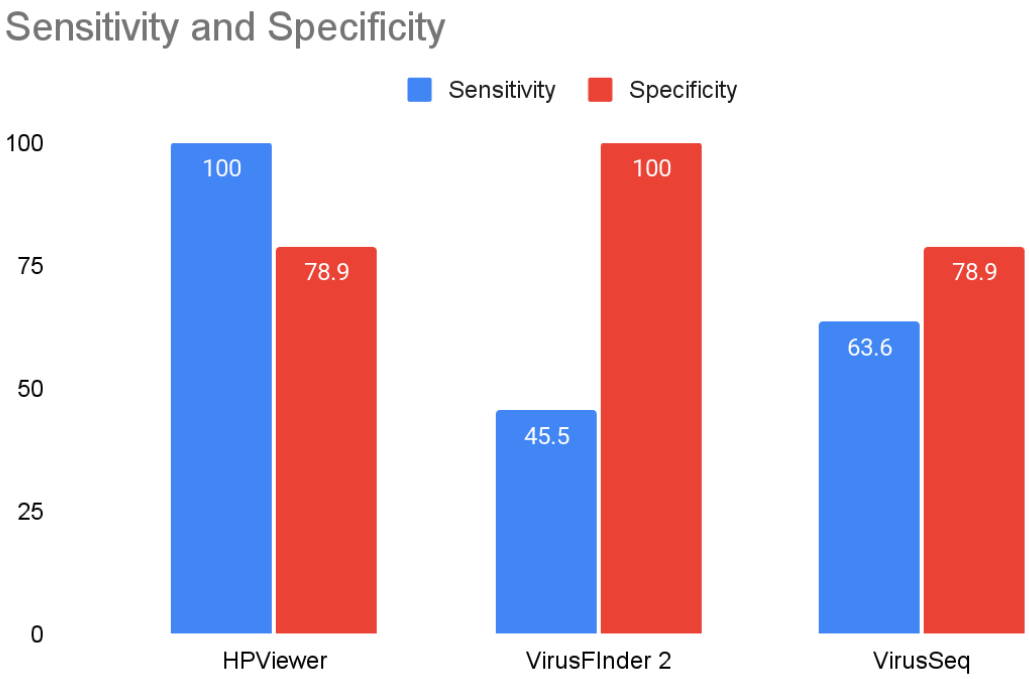

Figure 4: Sensitivity and Specificity of the Tested Tools (The red bar indicates the specificity score, while the blue bar indicates the sensitivity score for each tool.).

Detection of HPV type

Figure 5 shows the number of detections of each HPV type per tool from all the samples. It aims to determine the ability or inability of each tool to detect the HPV types. It shows that HPViewer had similar detection pattern as the reference, while VirusFinder 2 was limited in the number of HPV types it could detect. VirusSeq had more similarities with the reference and HPViewer than VirusFinder 2, indicating its ability to detect more HPV types. Table 3 shows the HPV types detected per sample by the tools; it shows that HPViewer had similar type detection as the reference, followed by VirusSeq. However, VirusFinder 2 was able to detect fewer number of HPV types compared to the others. In this study, 40 types were detected from the samples, including HPV 16 and 18, which are mostly linked to the disease (Aker et al., 2022).

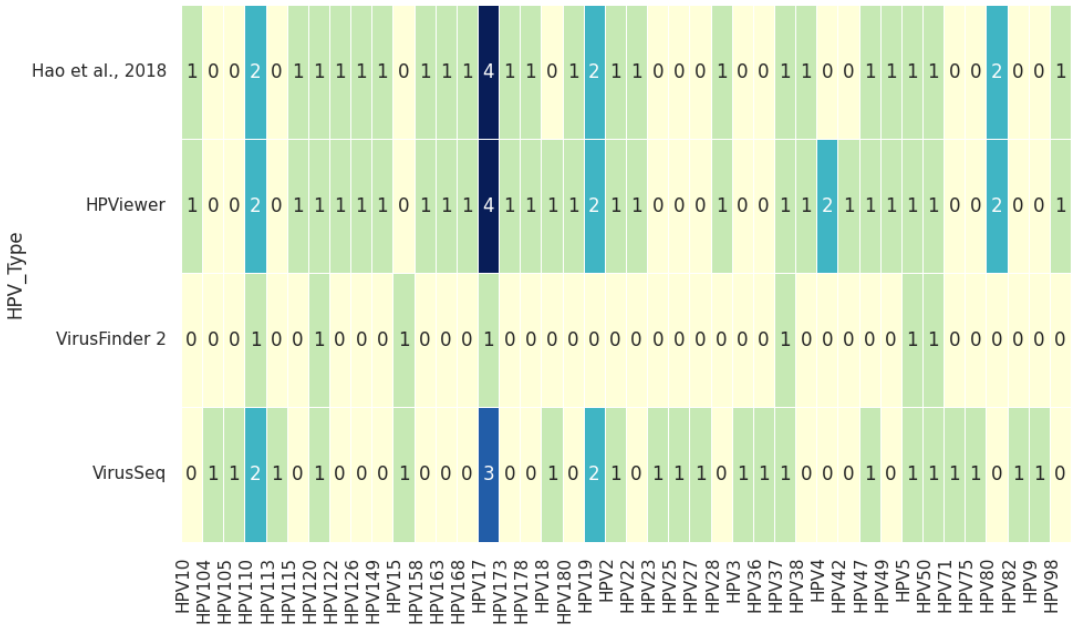

Figure 5: HPV type detection. The number in the cells indicates the count of detections of the HPV type per tool. Note that some samples had more than one HPV type; hence, the, total count per tool may be more than the number of samples.

Table 3: HPV type Detected per sample by the Tools

Id	Hao et al., 2018	HPViewer	VirusFinder 2	VirusSeq
SRS011132				
SRS011263				
SRS012663	HPV110, HPV17, HPV80	HPV110, HPV17, HPV80	HPV110	HPV110
SRS013876				
SRS015051	HPV17, HPV50	HPV17, HPV50	HPV17, HPV50	HPV17, HPV50
SRS015430				

SRS015640		HPV18		HPV18
SRS015752	HPV17, HPV178	HPV17, HPV178, HPV4		
SRS015996	HPV17, HPV173	HPV17, HPV173		
SRS016033		HPV2		HPV2, HPV27
SRS016105				
SRS016188				
SRS016292	HPV126, HPV180, HPV47, HPV80	HPV126, HPV180, HPV47, HPV80	HPV15, HPV37	HPV104, HPV105, HPV110, HPV113, HPV15, HPV17, HPV36, HPV37, HPV47, HPV75, HPV9
SRS016752	HPV122, HPV22, HPV49, HPV98	HPV122, HPV22, HPV49, HPV98		
SRS017044	HPV19, HPV28	HPV19, HPV28		HPV19, HPV25
SRS017244				
SRS017451				
SRS017697		HPV4		HPV71, HPV82
SRS018369				
SRS018463				
SRS018585				

SRS019067	HPV10, HPV115, HPV149	HPV10, HPV115, HPV149		
SRS019119				
SRS020386	HPV158, HPV36, HPV37, HPV5	HPV158, HPV37, HPV5	HPV5	HPV5
SRS021483	HPV110, HPV120, HPV38	HPV110, HPV120, HPV38	HPV120	HPV120, HPV23, HPV3
SRS023970				
SRS024567				
SRS044474	HPV163, HPV168, HPV19	HPV163, HPV168, HPV19		HPV19
SRS046973				
SRS054061		HPV42		HPV17

Discussion

This study was conducted with the primary goal of comparing the performance of three widely used bioinformatics tools in the context of detecting HPV (Human Papillomavirus) cervical cancer genomic data. HPV is a specific virus of interest, and for this analysis, both VirusFinder 2 and VirusSeq were optimized for HPV detection by exclusively utilizing the HPV sequences from the provided virus database. This optimized database was then used to generate reference indices and jump files for alignment. This study is of importance in highlighting an effective way of detecting HPV in biological samples, as HPV detection kits are unable to detect all types of HPV virus (Dubois et al., 2022).

While using VirusFinder 2, certain issues and bugs were encountered. These problems could be attributed to updates and upgrades in the third-party tools that VirusFinder 2 relies on. This outcome from the present study is in line with a previous report by Chen et al., (2019) who opined that older versions of the third-party tools had to be installed before the tool could be used. This highlights the potential challenges when incorporating third-party tools into bioinformatics pipelines, as updates may render them incompatible with previous versions. Remarkably, such issues were less prevalent with HPViewer, as it required fewer third-party tools in its operation. The relevance of the individual dependencies of the tools to their performance was buttressed by Waite et al., (2022), who showed that using different databases for the tools (for instance) results in different performance for the same tools. Hao et al. (2018) also highlight the importance of the components (algorithms, databases, third-party tools, etc.) of the tool to its performance. The authors while developing the tool used different types of database structures and obtained different sensitivity and specificity results eventually settling for a hybrid structure that maximizes both sensitivity and specificity. Wu et al., (2023) also buttressed that the algorithm and parameters used influence the ability of tools to detect viruses.

Human papillomavirus (HPV) comprises a diverse group of over 200 types, yet only a handful are notably associated with cervical cancer, earning them the classification of high-risk HPV (Okunade, 2019). Among these, HPV 16, HPV 18, HPV 31, HPV 33, HPV 45, HPV 52, and HPV 58 are the most prevalent culprits, collectively responsible for approximately 70% of cervical cancer cases. In particular, HPV 16 and HPV 18 loom as the most common, jointly contributing to roughly half of all cervical cancer diagnoses (HPV and Cancer - NCI, 2019).

In our comprehensive analysis, we scrutinized the performance of three distinct bioinformatics tools—HPViewer, VirusSeq, and VirusFinder 2—in detecting these high-risk HPV types, with a focus on HPV 18, an HPV type closely associated with cervical cancer. The results were revealing. HPViewer and VirusSeq exhibited a commendable ability to detect HPV 18, a significant marker of cervical cancer risk. This not only underscores their proficiency in identifying HPV from cervical cancer samples but also their suitability for research dedicated to understanding the role of HPV in cervical malignancies.

Conversely, VirusFinder 2, one of the tools evaluated, faced limitations in detecting any of the HPV types linked to cervical cancer. These findings highlight the need for a careful choice of bioinformatics tools when examining cervical cancer samples, especially when high-risk HPV types are the focus. The ability to accurately detect these HPV types can significantly impact our understanding of the disease and its management strategies.

The limitations in detecting certain HPV types using VirusFinder 2 might be linked to the absence of reference genomes for these specific types within its virus database. This could be a result of

incomplete or outdated information available at the time the tool was initially developed. It also emphasizes the effectiveness of the approach adopted by HPViewer in overcoming such limitations. The low detection ability of VirusFinder was also reported by (Waite et al., 2022). In their study, VirusFinder was limited in its ability to detect viruses from plant, vertebrate and invertebrate genomes while using the tool's default database, having a sensitivity of 44.42%, 36.54% and 46.29% respectively. These scores are similar to the sensitivity score obtained in this study for VirusFinder 2 at 45.5%. However, according to (Waite et al., 2022), there is a marked increase in sensitivity of VirusFinder 2 when detecting viruses from plants, vertebrates, and invertebrates. When the authors changed the database to modEPV, the new values were 91.49%, 80.83% and 86.68%, respectively. Hao et al., (2018) reported a sensitivity of 98.7 for HPViewer and this is close to the 100% sensitivity obtained in study for the tool. Wu et al., (2023) while benchmarking bioinformatic tools, including VirusFinder 2, also noted the tools low detection capability.

Moreover, the extended processing time observed with VirusSeq can be largely attributed to its utilization of the hashing system within the Mosaik aligner, upon which it depends. This was also pointed out by Chen et al., (2019), who stated that virus dependency on the Mosaik aligner made it require more time than the other tools to run. Despite efforts to reduce hash size, this tool still demanded more time and computing resources compared to the other options. In contrast, HPViewer demonstrated faster performance due to its specificity for HPV detection and the application of the Homology and Repeat Mask methods. In genomics, the Homology method involves comparing sequences of DNA or RNA to identify regions of similarity (Pearson 2012). In the context of HPViewer, this method leverages conserved sequences or regions within the HPV genome. By aligning the input sequences against these conserved regions, HPViewer can confidently identify and classify HPV with a higher degree of accuracy (Hao et al., 2018). This approach not only improves sensitivity but also ensures that potential false positives are minimized, contributing to the tool's overall specificity. Repetitive sequences can lead to misalignments and misinterpretations, impacting the accuracy of results. HPViewer addresses this challenge by applying the Repeat Mask method, which involves masking or filtering out these repetitive elements. This ensures that the analysis focuses on unique, informative segments of the genome, preventing ambiguity in HPV detection. The implementation of the Repeat Mask method contributes to the tool's robust performance by reducing the likelihood of false positives and enhancing overall precision (Hao et al., 2018). It tailored its approach to the characteristics of the virus, thereby expediting the analysis process. The speed of HPViewer was also highlighted in Hao et al., (2018), who stated that HPViewer had the fastest analysis speed among the different tools (Vipie, HPV detector) it was compared to. These findings underscore the importance of considering both the speed and accuracy of bioinformatic tools when making a selection.

HPViewer excels in terms of computational speed, making it suitable for researchers who prioritize rapid analysis. Notably, VirusSeq stood out in terms of resource requirements, consuming a substantial amount of computing resources and disk space, totaling 19 GB without considering the test sequences. Conversely, HPViewer exhibited efficiency by demanding minimal space and fewer resources, making it a favorable choice in terms of resource utilization.

A key observation from the study was that using the VirusSeq log files resulted in more positive results compared to using the default output. Researchers can leverage this insight to maximize their use of VirusSeq, thereby improving its performance.

These findings collectively emphasize the importance of considering the specific needs and constraints of a research project when selecting a bioinformatics tool. The trade-offs between speed, accuracy, and resource consumption must be carefully weighed. Researchers should be aware of potential issues related to third-party tools and stay updated on any available patches or solutions to ensure smooth workflow integration. Additionally, the research highlights the need for regular updates and expansions of virus databases to enhance the detection capabilities of bioinformatic tools. Future research in this area should focus on refining these tools, expanding their reference databases, and improving documentation and user-friendliness to cater to a wider scientific community effectively.

Potential Scientific Contribution

The present study is geared towards comparing the NGS tools. The comparison of some novel NGS tools, such as HPViewer, VirusSeq, and VisurFinder 2, among others, will affirm the sensitivity and accuracy of their roles in bioinformatic research. The benefit of the NGS-based approach in bioinformatic research and development is that they allow for the identification of various techniques that could serve as cost-effective and more accurate alternatives in scientific research especially in bioinformatics. Thus, it will open new horizons for further research references.

This research contributes to the field by providing a comprehensive evaluation of popular bioinformatics tools for HPV detection, allowing researchers and clinicians to make informed decisions based on their specific research objectives, resources, and priorities. It highlights the trade-offs between sensitivity, specificity, and resource requirements and underlines the importance of database completeness and regular tool updates. In the rapidly evolving landscape of bioinformatics, this work encourages the development of more user-friendly tools, detailed documentation, and comprehensive databases to facilitate and improve the efficiency of HPV detection.

NGS in recent times has transformed the landscape of scientific research with a wide range of applications in whole-genome sequencing, transcriptome profiling, metagenomics, and disease diagnosis and surveillance, among other applications (Gargis, Kalman and Lubin, 2016). For example, NGS has recently been adopted in many scientific fields of study. This can be seen in the use of NGS tools to detect HPV in cervical cancerous cells (Yan et al., 2019). The applications of the outcome of the present study coupled with recent advancements in different HPV detecting tools on cervical cancer and the inference that will be made from that will lead to the combination of different strengths of the tools to build a pipeline that combines them in some way. This has the potential to either complement existing scientific methods or even replace them with rapid bioinformatics pipelines.

Global Health Implications

Cervical cancer's disproportionate impact on low and middle-income countries (LMICs), representing about 87% of global cases, underscores the urgency for enhanced detection methods in resource-limited regions (Torre et al., 2012). Despite the efficacy of HPV vaccination in preventing up to 90% of cervical cancer cases, limited uptake in many LMICs due to cost and access barriers persists. Advances in early detection could prove crucial, especially in unvaccinated populations. While low-cost visual inspection methods like VIA offer cervical screening options for LMICs, their limited specificity necessitates more advanced yet cost-effective screening tools for timely intervention (Hull et al., 2020). Local manufacturing or licensing arrangements for diagnostic tests and use of NGS and bioinformatic techniques could not only create business opportunities but also enhance accessibility in developing regions, fostering economic growth in health and biotech sectors. Given the intricate link between cervical cancer and HIV co-infection, improved detection of precancerous lesions aligns with HIV testing and treatment initiatives in regions like sub-Saharan Africa, showcasing broader synergies in public health.

Gender Inequality

Cervical cancer exerts a heavier toll on women in low and middle-income countries (LMIC), with a staggering 85% of both cases and fatalities concentrated in these regions (Randall and Ghebre, 2016). The disparity is stark, as the death rate from cervical cancer in LMICs is 18 times higher than in developed nations. Globally, cervical cancer stands as the second most prevalent cause of cancer-related deaths among women. This burden is markedly more pronounced in LLMICs where both incidence and mortality rates surpass those in developed nations (LaVigne et al., 2017). The limited resources allocated to cervical cancer prevention, screening, and treatment in LMIC exacerbate the challenges faced by women, creating a formidable health inequity. Despite concerted efforts, the fight against cervical cancer in LMIC grapples with obstacles such as deficient infrastructure, restricted

access to screening, treatment, and preventive HPV vaccines, along with a shortage of skilled professionals and training opportunities (LaVigne et al., 2017).

Ethical Considerations

In conducting this research, strict adherence to ethical guidelines has been a priority. The study revolves around the detection of Human Papillomavirus (HPV) in cervical cancer using secondary data from the Human Microbiome Project (HMP). All ethical standards, including data privacy, informed consent, and participant rights, have been diligently observed.

Given the sensitivity of cervical cancer, our foremost ethical commitment is to ensure the accuracy of diagnostic information. Misinterpretations or inaccuracies in the bioinformatics tools could have serious implications, potentially leading to misdiagnosis and inappropriate treatments. This ethical imperative aligns with the responsibility to prioritize patient well-being.

Transparency is a key ethical principle in scientific practice. The research provides a comprehensive understanding of the evaluated tools, their limitations, and the specific conditions under which they were tested. This commitment to transparency facilitates the reproducibility of results, allowing fellow researchers to validate findings and contribute to the collective knowledge in the field.

Respecting participant rights and data privacy is integral. The use of secondary data from the HMP is guided by ethical norms, ensuring that the data is used for its intended purpose and with due consideration for participant confidentiality. Upholding the highest standards of privacy and confidentiality is a non-negotiable ethical commitment.

The study also recognizes the societal implications of its outcomes, particularly in the context of HPV detection in cervical cancer. While adhering to ethical guidelines, the research aims to contribute responsibly to public health knowledge and decision-making.

In summary, the research has diligently followed ethical guidelines, emphasizing accuracy in diagnosis, transparency in reporting, and a steadfast commitment to participant rights and privacy. The use of secondary data from the HMP has been conducted with the utmost respect for ethical considerations, ensuring the responsible advancement of knowledge in the field of cervical cancer detection.

Conclusion

In conclusion, this study contributes to the ongoing efforts in the field of bioinformatics by offering a comparative analysis of tools for HPV detection, enabling researchers to make informed choices for their specific research needs. The findings presented here have implications for those engaged in HPV-related research and emphasize the necessity of optimizing existing tools, expanding reference databases, and enhancing the usability of bioinformatics resources.
The study revealed that HPViewer had better performance in terms of speed and sensitivity than the other tools, while VirusFinder 2 had better specificity result, although this came at the cost of speed. VirusSeq had a more balanced performance across the parameters measured

The study underscores the importance of understanding the trade-offs between speed, accuracy, and resource consumption when choosing a bioinformatics tool for HPV detection. While each tool exhibits unique strengths and limitations, they collectively provide valuable options for HPV detection, catering to the diverse requirements of researchers in the field of viral genomics.

With the field of bioinformatics continually evolving, there is ample room for future research and development aimed at enhancing the performance and user-friendliness of these tools. As new data emerges and our understanding of viral genomics deepens, the tools and methodologies employed must adapt to meet the evolving demands of this critical research area.

Limitations of Study

It is important to note the limitations of this study which are the limited number of samples used. A larger number could have made for a more encompassing result.

References

Arroyo Mühr, L.S. *et al.* (2020) 'Deep sequencing detects human papillomavirus (HPV) in cervical cancers negative for HPV by PCR', *British Journal of Cancer*, 123(12), pp. 1790–1795. Available at: https://doi.org/10.1038/s41416-020-01111-0.

Augustin, J.G. *et al.* (2020) 'HPV Detection in Head and Neck Squamous Cell Carcinomas: What Is the Issue?', *Frontiers in Oncology*, 10(September), pp. 1–13. Available at: https://doi.org/10.3389/fonc.2020.01751.

Bettoni, F. *et al.* (2017) 'A straightforward assay to evaluate DNA integrity and optimize next-generation sequencing for clinical diagnosis in oncology', *Experimental and Molecular Pathology*, 103(3), pp. 294–299. Available at: https://doi.org/10.1016/j.yexmp.2017.11.011.

Chandrani, P. *et al.* (2015a) 'NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome', *British Journal of Cancer*, 112(12), pp. 1958–1965. Available at: https://doi.org/10.1038/bjc.2015.121.

Chandrani, P. *et al.* (2015b) 'NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome', *British Journal of Cancer*, 112(12), pp. 1958–1965. Available at: https://doi.org/10.1038/bjc.2015.121.

Chandrani, P. *et al.* (2015c) 'NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome', *British Journal of Cancer*, 112(12), pp. 1958–1965. Available at: https://doi.org/10.1038/bjc.2015.121.

Chen, Y. *et al.* (2013) 'VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue', *Bioinformatics*, 29(2), pp. 266–267. Available at: https://doi.org/10.1093/bioinformatics/bts665.

Dubois, B. *et al.* (2022) 'A detailed workflow to develop QIIME2-formatted reference databases for taxonomic analysis of DNA metabarcoding data', *BMC Genomic Data*, 23(1), pp. 1–14. Available at: https://doi.org/10.1186/s12863-022-01067-5.

Fraser, T.A. *et al.* (2019) 'Quantitative real-time PCR assay for the rapid identification of the intrinsically multidrug-resistant bacterial pathogen Stenotrophomonas maltophilia', *Microbial Genomics*. Available at: https://doi.org/10.1099/mgen.0.000307.

Gargis, A.S., Kalman, L. and Lubin, M. (2016) 'Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories', 54(12), pp. 2857–2865. Available at:

https://doi.org/10.1128/JCM.00949-16.Editor.

Gates, A. *et al.* (2021) 'Screening for the prevention and early detection of cervical cancer: protocol for systematic reviews to inform Canadian recommendations', *Systematic Reviews*, 10(1). Available at: https://doi.org/10.1186/s13643-020-01538-9.

Goswami, R.S. (2016) 'PCR techniques in next-generation sequencing', *Methods in Molecular Biology*, 1392, pp. 143–151. Available at: https://doi.org/10.1007/978-1-4939-3360-0_13.

Hao, Y. *et al.* (2018) 'HPViewer: Sensitive and specific genotyping of human papillomavirus in metagenomic DNA', *Bioinformatics*, 34(12), pp. 1986–1995. Available at: https://doi.org/10.1093/bioinformatics/bty037.

Hathaway, J.O.N.K. (2012) 'HPV : Diagnosis , Prevention , and Treatment', 55(3), pp. 671–680.

He, W. *et al.* (2021) 'Research Integrated Network of Systems (RINS): A virtual data warehouse for the acceleration of translational research', *Journal of the American Medical Informatics Association*, 28(7), pp. 1440–1450. Available at: https://doi.org/10.1093/jamia/ocab023.

Helene, C. and Francois, R. (2013) 'Introduction to bioinformatics', *Functional Plant Genomics*, (June), pp. 53–55. Available at: https://doi.org/10.1201/b13091-1.

Hu, Z. *et al.* (2015) 'Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism', *Nature Genetics*, 47(2), pp. 158–163. Available at: https://doi.org/10.1038/ng.3178.

Hu, Z. and Ma, D. (2018) 'The precision prevention and therapy of HPV-related cervical cancer: new concepts and clinical implications', *Cancer Medicine*, pp. 5217–5236. Available at: https://doi.org/10.1002/cam4.1501.

Hull, R., Mbele, M., Makhafola, T., Hicks, C., Wang, S. M., Reis, R. M., Mehrotra, R., Mkhize-Kwitshana, Z., Kibiki, G., Bates, D. O., & Dlamini, Z. (2020). Cervical cancer in low and middle-income countries. Oncology letters, 20(3), 2058–2074. https://doi.org/10.3892/ol.2020.11754

Khan, A. *et al.* (2019) 'Detection of human papillomavirus in cases of head and neck squamous cell carcinoma by RNA-seq and VirTect', *Molecular Oncology*, 13(4), pp. 829–839. Available at: https://doi.org/10.1002/1878-0261.12435.

Lanigan, T.M., Kopera, H.C. and Saunders, T.L. (2020) 'Principles of genetic engineering', *Genes*. Available at: https://doi.org/10.3390/genes11030291.

LaVigne, A. W., Triedman, S. A., Randall, T. C., Trimble, E. L., & Viswanathan, A. N. (2017). Cervical

cancer in low and middle income countries: Addressing barriers to radiotherapy delivery. Gynecologic oncology reports, 22, 16–20. https://doi.org/10.1016/j.gore.2017.08.004

Lee, J.Y. *et al.* (2020) 'Next Generation Sequencing Assay for Detection of Circulating HPV DNA (cHPV-DNA) in Patients Undergoing Radical (Chemo) Radiotherapy in Anal Squamous Cell Carcinoma (ASCC)', 10(April), pp. 1–7. Available at: https://doi.org/10.3389/fonc.2020.00505.

Li, J.W. *et al.* (2013) 'ViralFusionSeq: Accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution', *Bioinformatics*, 29(5), pp. 649–651. Available at: https://doi.org/10.1093/bioinformatics/btt011.

Li, W. *et al.* (2013) 'HIVID: An efficient method to detect HBV integration using low coverage sequencing', *Genomics*, 102(4), pp. 338–344. Available at: https://doi.org/10.1016/j.ygeno.2013.07.002.

Markowetz, F. (2017) 'All biology is computational biology', *PLoS Biology*, 15(3), pp. 4–7. Available at: https://doi.org/10.1371/journal.pbio.2002050.

Moreau, F. *et al.* (2013) 'Detection and genotyping of human papillomavirus by real-time PCR assay', *Journal of Clinical Virology*, 56(3), pp. 328–333. Available at: https://doi.org/10.1016/j.jcv.2012.11.003.

Mulder, N.J. *et al.* (2018) 'HHS Public Access', 12(2), pp. 91–98. Available at: https://doi.org/10.1016/j.gheart.2017.01.005.Development.

Naeem, R., Rashid, M. and Pain, A. (2013) 'READSCAN: A fast and scalable pathogen discovery program with accurate genome relative abundance estimation', *Bioinformatics*, 29(3), pp. 391–392. Available at: https://doi.org/10.1093/bioinformatics/bts684.

Nilyanimit, P. *et al.* (2018) 'Comparison of four human papillomavirus genotyping methods: Next-generation sequencing, INNO-LiPA, electrochemical DNA Chip, and nested-PCR', *Annals of Laboratory Medicine*, 38(2), pp. 139–146. Available at: https://doi.org/10.3343/alm.2018.38.2.139.

Pimple, S. and Mishra, G. (2022) 'Cancer cervix: Epidemiology and disease burden', *CytoJournal*. Available at: https://doi.org/10.25259/CMAS_03_02_2021.

Pongor, S. and Landsman, D. (2015) 'Bioinformatics and developing word', *Biotechnology and development monitor*, 40, pp. 1–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25960606%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4422384.

Qiu, L. *et al.* (2022) 'Characterization of the Genomic Landscape in Cervical Cancer by Next Generation Sequencing', *Genes*, 13(2). Available at: https://doi.org/10.3390/genes13020287.

Randall, T. C., & Ghebre, R. (2016). Challenges in Prevention and Care Delivery for Women with Cervical Cancer in Sub-Saharan Africa. Frontiers in oncology, 6, 160. https://doi.org/10.3389/fonc.2016.00160

Sendagorta-cudós, E. and Burgos-cibrián, J. (2019) 'Infecciones genitales por el virus del papiloma humano', *Enfermedades Infecciosas y Microbiología Clínica*, 37(5), pp. 324–334. Available at: https://doi.org/10.1016/j.eimc.2019.01.010.

Shen-gunther, J. *et al.* (2021) 'HPV DeepSeq : An Ultra-Fast Method of NGS Data Analysis and Visualization Using Automated Workflows and a Customized Papillomavirus Database in CLC Genomics Workbench'.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. CA: a cancer journal for clinicians, 65(2), 87–108. https://doi.org/10.3322/caac.21262

Ure, A., Mukhedkar, D. and Arroyo Mühr, L.S. (2022) 'Using HPV-meta for human papillomavirus RNA quality detection', *Scientific Reports*, 12(1), pp. 1–8. Available at: https://doi.org/10.1038/s41598-022-17318-5.

Visser, M., Burger, J.T. and Maree, H.J. (2016) 'Targeted virus detection in next-generation sequencing data using an automated e-probe based approach', *Virology*, 495, pp. 122–128. Available at: https://doi.org/10.1016/j.virol.2016.05.008.

Wang, Q. (2015) 'VirusFinder2.0-manual', (version 2).

Wang, Q., Jia, P. and Zhao, Z. (2013) 'VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data', *PLoS ONE*, 8(5), pp. 1–5. Available at: https://doi.org/10.1371/journal.pone.0064465.

Yan, B. *et al.* (2019) 'DisV-HPV16, versatile and powerful software to detect HPV in RNA sequencing data', *BMC Infectious Diseases*, 19(1), pp. 1–7. Available at: https://doi.org/10.1186/s12879-019-4123-z.

Appendix

Table 1: Data characteristics

SRS ID	Body Site	Reads File Location	Reads File Size
SRS011132	anterior_nares	/data/Illumina/anterior_nares/SRS011132.tar.bz2	30874163
SRS011263	anterior_nares	/data/Illumina/anterior_nares/SRS011263.tar.bz2	30630950
SRS012663	anterior_nares	/data/Illumina/anterior_nares/SRS012663.tar.bz2	105667144
SRS013876	anterior_nares	/data/Illumina/anterior_nares/SRS013876.tar.bz2	177409068
SRS015051	anterior_nares	/data/Illumina/anterior_nares/SRS015051.tar.bz2	36481771
SRS015430	anterior_nares	/data/Illumina/anterior_nares/SRS015430.tar.bz2	35575892
SRS015640	anterior_nares	/data/Illumina/anterior_nares/SRS015640.tar.bz2	18415433
SRS015752	anterior_nares	/data/Illumina/anterior_nares/SRS015752.tar.bz2	77260801
SRS015996	anterior_nares	/data/Illumina/anterior_nares/SRS015996.tar.bz2	365463394
SRS016033	anterior_nares	/data/Illumina/anterior_nares/SRS016033.tar.bz2	25555143
SRS016105	anterior_nares	/data/Illumina/anterior_nares/SRS016105.tar.bz2	72239353
SRS016188	anterior_nares	/data/Illumina/anterior_nares/SRS016188.tar.bz2	35536436
SRS016292	anterior_nares	/data/Illumina/anterior_nares/SRS016292.tar.bz2	85435039
SRS016752	anterior_nares	/data/Illumina/anterior_nares/SRS016752.tar.bz2	85943169
SRS017044	anterior_nares	/data/Illumina/anterior_nares/SRS017044.tar.bz2	91197174

SRS017244	anterior_nares	/data/Illumina/anterior_nares/SRS017244.tar.bz2	24687838
SRS017451	anterior_nares	/data/Illumina/anterior_nares/SRS017451.tar.bz2	78565926
SRS017697	anterior_nares	/data/Illumina/anterior_nares/SRS017697.tar.bz2	48153493
SRS018369	anterior_nares	/data/Illumina/anterior_nares/SRS018369.tar.bz2	25614289
SRS018463	anterior_nares	/data/Illumina/anterior_nares/SRS018463.tar.bz2	35647099
SRS018585	anterior_nares	/data/Illumina/anterior_nares/SRS018585.tar.bz2	17652230
SRS019067	anterior_nares	/data/Illumina/anterior_nares/SRS019067.tar.bz2	43457527
SRS019119	anterior_nares	/data/Illumina/anterior_nares/SRS019119.tar.bz2	51365975
SRS020386	anterior_nares	/data/Illumina/anterior_nares/SRS020386.tar.bz2	251423053
SRS021483	anterior_nares	/data/Illumina/anterior_nares/SRS021483.tar.bz2	40815560
SRS023970	anterior_nares	/data/Illumina/anterior_nares/SRS023970.tar.bz2	38143134
SRS024567	anterior_nares	/data/Illumina/anterior_nares/SRS024567.tar.bz2	25654232
SRS044474	anterior_nares	/data/Illumina/anterior_nares/SRS044474.tar.bz2	241202975
SRS046973	anterior_nares	/data/Illumina/anterior_nares/SRS046973.tar.bz2	54134873