viruses

MDPI

*Article*

# Human Papillomavirus Detection by Whole-Genome Next-Generation Sequencing: Importance of Validation and Quality Assurance Procedures

**Laila Sara Arroyo Mühr [1,†], Daniel Guerendiain [2,3,†], Kate Cuschieri [2,‡] and Karin Sundström [4,*,‡]**

[1] International HPV Reference Center, Department of Laboratory Medicine, Karolinska Institutet, SE-141 86 Stockholm, Sweden; sara.arroyo.muhr@ki.se
[2] Scottish Human Papillomavirus Reference Laboratory (SHPVRL), Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK; Kate.Cuschieri@nhslothian.scot.nhs.uk
[3] School of Medicine, University of St Andrews, St Andrews KY16 9TF, UK; dgr7@st-andrews.ac.uk
[4] Department of Laboratory Medicine, Karolinska Institutet, SE-141 86 Stockholm, Sweden
* Correspondence: Karin.sundstrom@ki.se
† Equal contribution.
‡ Equal contribution.

**Abstract:** Next-generation sequencing (NGS) yields powerful opportunities for studying human papillomavirus (HPV) genomics for applications in epidemiology, public health, and clinical diagnostics. HPV genotypes, variants, and point mutations can be investigated in clinical materials and described in previously unprecedented detail. However, both the NGS laboratory analysis and bioinformatical approach require numerous steps and checks to ensure robust interpretation of results. Here, we provide a step-by-step review of recommendations for validation and quality assurance procedures of each step in the typical NGS workflow, with a focus on whole-genome sequencing approaches. The use of directed pilots and protocols to ensure optimization of sequencing data yield, followed by curated bioinformatical procedures, is particularly emphasized. Finally, the storage and sharing of data sets are discussed. The development of international standards for quality assurance should be a goal for the HPV NGS community, similar to what has been developed for other areas of sequencing efforts including microbiology and molecular pathology. We thus propose that it is time for NGS to be included in the global efforts on quality assurance and improvement of HPV-based testing and diagnostics.

**Keywords:** human papillomavirus; HPV; next-generation sequencing; NGS; whole-genome sequencing; WGS; deep sequencing



**Citation:** Arroyo Mühr, L.S.; Guerendiain, D.; Cuschieri, K.; Sundström, K. Human Papillomavirus Detection by Whole-Genome Next-Generation Sequencing: Importance of Validation and Quality Assurance Procedures. *Viruses* **2021**, *13*, 1323. https://doi.org/10.3390/v13071323

Academic Editors: Lisa Mirabello and Meredith Yeager

Received: 6 May 2021
Accepted: 18 June 2021
Published: 8 July 2021



## 1. Introduction

Tests for the detection of human papillomavirus (HPV) infection in humans have evolved dramatically over the last decades. Initial low-throughput hybridization/blotting techniques prefaced broad-spectrum signal amplification assays, which were then replaced by rapid high-throughput target-amplification assays involving quantitative polymerase chain reaction (qPCR). The latter tests are capable of detecting individual HPV-genotypes and have become the mainstay of HPV-based screening and clinical testing [1]. Arguably, the next "age" of HPV testing should involve going beyond simply detecting the presence or absence of HPV but, rather, providing more detailed insight into the likely course and clinical consequences of HPV infection.

Next-generation sequencing (NGS) of human or microbial genetic material is being applied increasingly in laboratory contexts, to facilitate research, population-based epidemiology, and recently, personalized patient diagnostics. NGS can be used as a highly sensitive method for HPV detection due to its ability to detect types at low copy number (even within multiple infections), novel types, and/or known types that are distantly related to

primers/probes which may escape detection using standard molecular approaches [2–4]. When employing whole-genome NGS, which covers the entire genome and not only exomes or targeted regions, we conveniently allow high-accuracy determination of sequences below the phylogenetic level of genotype, i.e., variants and subvariants of HPV [3–8]. Indeed, in recent years, various studies have utilized NGS approaches to generate detailed insights into potential disease-related mechanisms of HPV. NGS has shown that certain sublineages of HPV are associated with a higher risk of cancer [9–11], and its high sensitivity also allows the attributable fraction of cervical cancer associated with HPV to be determined with greater precision compared to traditional PCR techniques [12,13]. Further, NGS has identified certain single-nucleotide polymorphisms (SNPs) associated with a higher likelihood of viral persistence [14] and the key role of HPV *E7* gene conservation in cervical cancer development [15].

Although NGS has been used in HPV research for some years with different applications, as described above, NGS in clinical diagnosis is not yet extensively used. Lack of standardization and quality guidelines, as well as expense and requirement for ancillary laboratory infrastructure, may have slowed down the adoption of this technology in clinical laboratories. However, as is discussed later, NGS use could ultimately improve diagnosis and management of patients with HPV-driven lesions. This includes utilities ranging from a more sensitive detection of HPV to detection of true viral persistence [13], identification of risk according to HPV sublineage [9–11], and detection of circulating HPV DNA in patients who have received cancer treatment [16].

NGS is a technology based on massively parallel sequencing or "deep sequencing" of nucleic acid sequences. Nucleic acid sequences are fragmented with each fragment being amplified and sequenced multiple times, providing a depth of information which can deliver accurate data at the nucleotide level. NGS can sequence the entire genome of HPV or be limited to specific areas of interest [11,14].

There are several general approaches to sequencing, depending on the size (i.e., length) of nucleotide reads obtained and detection method employed. Illumina and IonTorrent instruments obtain reads that are approximately 250 base pair (bp) long, whereas Oxford Nanopore MinIon and PacBio obtain longer reads, potentially exceeding 10,000 bp in length [17].

The whole NGS process requires several steps, which include initial sample identification, processing (nucleic acid extraction), viral enrichment (optional), library preparation, sequencing, and bioinformatic analysis of the raw data (Figure 1). While some of these steps are consistent with general requirements for molecular detection (i.e., sample extraction), the described downstream aspects arguably require an additional set of skills and analyses. Given the multistep nature of the process and the generation of large amounts of detailed data generated, it is essential that, where possible, standardization and quality checks to support consistency and integrity of data outputs are considered and implemented. For other applications, including those relevant to bacteriology and molecular pathology, quality guidelines have already been developed [18–20]. However, these are still lacking for the HPV field. Our chief aim is to provide a comprehensive starting point to widen perspectives and give practical advice for those who are new(er) to this topic, thus lowering barriers to introduce NGS specifically in HPV-based research and clinical applications.

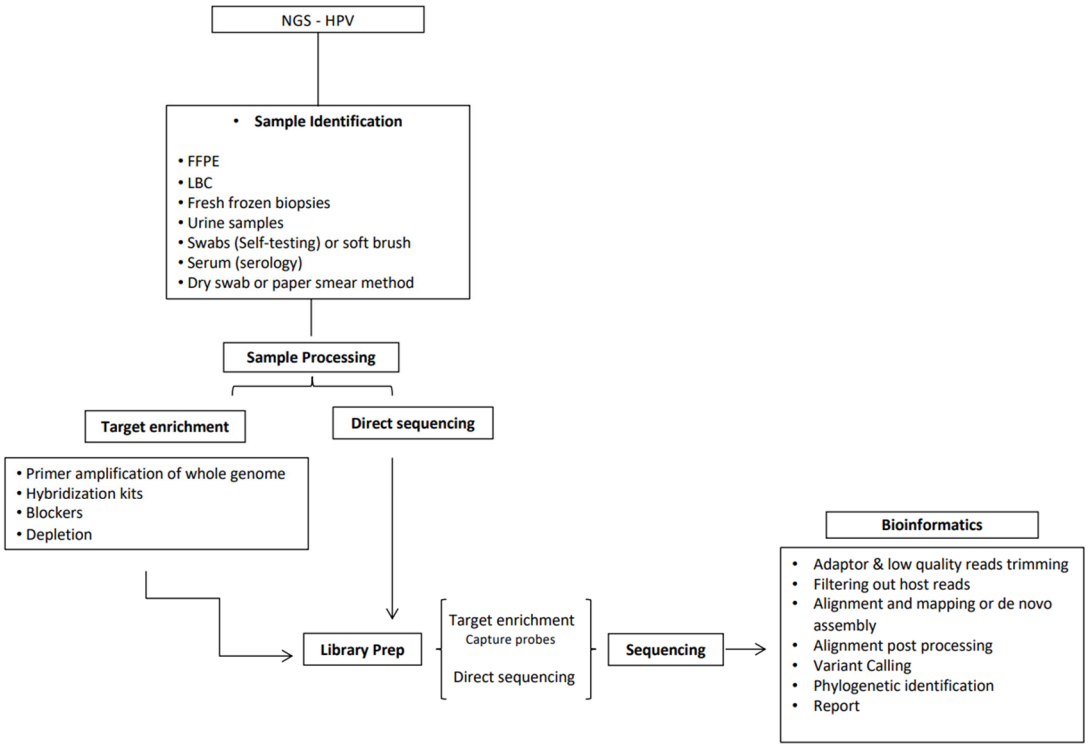


**Figure 1.** Next-generation sequencing (NGS) for human papillomavirus (HPV) detection and characterization process steps, from sample preparation to data analysis, with focus on whole-genome sequencing (WGS).

We are active in the field of HPV testing at the International HPV Reference Laboratory and the Scottish HPV Reference Laboratory (SHPVRL). Both entities were established in 2008, under the auspices of WHO (LabNet) and through the National Health Service of Scotland respectively. As such, we are committed to the evaluation and application of new technologies, including NGS technology, specifically to support the prevention and management of human cancers associated with HPV. In the present work we discuss the key stages of HPV-specific applications utilizing NGS and offer practical suggestions for quality-assurance procedures to support these stages, focusing on HPV whole-genome sequencing approaches. We believe that our experience may facilitate the implementation of NGS in laboratory settings so that it can play an increasing role in research, epidemiology, and importantly clinical testing. With respect to the latter, this is an important consideration, given the increased incorporation of NGS systems into routine departments of laboratory medicine, partly because of national strategies designed to harness the benefits of genomic medicine at the patient level in an agile but comprehensive way [21]. NGS is also consistent with the concept and aspirations of precision medicine, defined as an "approach for disease treatment and prevention that takes into account variability in genes, environment and lifestyle for each patient" [22].

## 2. Materials and Methods

### *2.1. NGS-Process Step 1: Laboratory Procedures*

#### 2.1.1. Pre-Analytical Sample Processing

All specimens used for the identification of HPV by existing molecular tests can in theory be used for NGS. However, various biospecimen types may present their own unique challenges. Liquid-based cytology samples, swab samples, or fresh frozen biopsies may typically be readily applied on an NGS platform. However, cells and tissue derived from formalin-fixed paraffin embedded (FFPE) material are more likely to be associated with fragmented nucleic acid and crosslinks between intracellular macromolecules such as proteins and DNA. Fragmentation can be a rate-limiting factor in approaches that demand

longer amplicons, which is why it is useful to note that researchers have successfully used shorter amplicons when working with FFPE DNA using NGS technologies [23–26].

#### 2.1.2. Nucleic Acid Extraction Method

Ideally, extraction methods should be assessed with a pilot panel of samples before embarking on a large project, in order to determine the quality and suitability of the specific extract for downstream NGS. Where possible, we recommend investigators evaluate at least two different extraction technologies to maximize nucleic acid yield. The quality, quantity, and fragment length prior to the library preparation must be determined, as different library preparations may require different nucleic acid input, quality, and length recommendations for library success. For Illumina (San Diego, CA, USA) DNA libraries for example, most of the protocols are optimized for 1 ng of input. Assessing the DNA purity is needed to ensure that the extract does not contain possible contaminants (EDTA, phenol, and ethanol) which can result in assay failure. UV absorbance is a common method used for assessing the purity of a DNA sample, and protocols generally define the "pure"/acceptable range as having an absorbance ratio values of 1.8–2.0 [27,28]. Highly fragmented nucleic acid extraction can lead to missing regions and sections with a low number of reads (low coverage), and the analysis may fail if subjected to enrichment-based amplification methods (see further below) [29,30]. Fragment length can be analyzed using a qPCR with specific length amplicons (same as target amplicon size if PCR-based enrichment is used) or through gel electrophoresis. As described above, a sample being highly fragmented prior to library preparation does not definitively preclude it from sequencing; however, knowing fragment length informs downstream options (e.g., if material is highly fragmented, operators can opt to skip fragmentation steps within the library preparation). HPV positive (with known HPV sequence) and negative internal control samples are necessary to demonstrate that nucleic material has been correctly extracted during the extraction process.

#### 2.1.3. Specimen Enrichment Approach

While HPV-positive clinical samples contain HPV nucleic acid, they are naturally dominated by nucleic acid from non-HPV sources (i.e., human and other-microbiome). If NGS is performed directly on the extract, without a targeted approach, the HPV content is at a relatively low proportion vs. the total nucleic acid sample. As the human genome length is 3 billion base pairs vs. ~8000 base pairs for HPV, the relative proportion of human nucleic acids extracted is +200,000%.

Sequencing studies reveal that viruses typically represent less than 1% of the total genomic material detected in a human specimen [7], and therefore, detection of any virus by NGS formerly required subjecting specimens to either (a) host genome depletion or (b) viral enrichment, first. Different approaches to increase the viral component, include low-speed or high-speed gradient centrifugation, separation of long chromosomal DNA, digestion of nucleic acids not protected by virions (e.g., nuclease treatment), filtration to remove bacterial and host cells, or targeted sequence capture [31–36]. Each of these procedures may bias against detection of some viruses; therefore, pilot studies to validate accuracy and reproducibility of the method for the investigator's specific purpose are necessary. A variety of methods have been described to enhance the HPV content of a sample which are described below.

(A). Depletion Protocols Using Saponin- or Lysis-Based Methods

Saponin is a non-ionic surfactant that depletes the human genome affecting the pathogen-human DNA ratio [37,38]. MolYsis (Molzym, Germany) is a commercial product which works through selective lysis of host cells and associated degradation of released host DNA. Both products thus reduce the amount of host nucleic acid, enrichening the HPV DNA (or other desired bacterial, fungal, or viral DNA) while simultaneously removing potential PCR inhibitors. However, replicating/intracellular viruses are also depleted with these methods, and potential loss of viral signaling can occur, especially for viruses (such as HPV) that integrate in the human genome.

A particular NGS application related to HPV is the detection of viral integration sites into the human genome. When performing WGS, authors have reported that tumors had either a small or a very large deletion in the viral genome and discovered that these deletions were the result of either HPV integration into the human genome or HPV–HPV sequence junctions [39]. It has further been reported that at least 83% [40,41] of cervical cancers with HPV infection have HPV integration, which can occur at any chromosome but more frequently at certain fragile sites [42]. HPV integration can significantly increase related gene expression and has been associated with a worse survival rate (compared to those with episomal HPV) [43–46]. HPV integration status may therefore have promise as a biomarker for risk stratification [47–49], including the monitoring of treatment and therapy. Different methods have been used to study HPV integration sites (e.g., amplification of HPV oncogene transcripts and detection of integrated HPV sequences by ligation-mediated PCR). With the development of NGS, whole-genome sequencing has been used for virus integration sites detection [50,51]. However, it requires large amounts of sequencing data, and thus is not applicable in clinical usage, which requires fast and accurate results. To date, the development of new NGS methods for HPV integration detection with high accuracy and prompt reporting capacity is ongoing [52]. However, the best way currently to detect integration sites is reached by using probe-captured sequencing methods (see section "Enrichment protocols" below). After enrichment of virus genomic material, the fusion fragment of human and HPV sequence is isolated and further sequenced by NGS.

(B). Enrichment Protocols

In addition to depletion protocols, the two most common enrichment protocols for HPV are PCR enrichment in the absence of chemical components and using capture protocols. PCR enrichment involves performing a PCR reaction prior to library preparation. For WGS, this technique may include amplification of the whole genome of the target virus via overlapping primers covering the entire genome (other techniques may not require amplification of the whole genome). A disadvantage of this technique is that there is a need to know the target that is to be sequenced a priori—thus, the method is not valid for detection of novel or nontargeted HPVs. Additionally, in the case of fragmented material, the number of primers required can be high, which requires a structured a priori design approach, and the primers may vary in their affinity for the separate regions. Primer-dimer formation can also limit efficient target amplification, and therefore, pilot studies are needed to validate the protocol for each specific purpose [9,53].

To overcome these bottlenecks, unbiased amplification (not based on PCR) has been commonly used for viral enrichment, as it amplifies all DNA material present in the sample. Multiple displacement amplification (MDA) is the gold standard method for non-PCR-based amplification techniques, where the reaction is based on annealing random hexamer primers to the DNA template [54]. MDA provides an effective way of amplifying minimal quantities of DNA, but there exist biases associated with this technology. Chimera formation, preferential amplification of circular single stranded DNA, and nonuniform amplification of linear genomes have been documented [55,56]. Authors have quantified the amount of amplification of both human DNA and HPV DNA by adding 20 copies/μL of HPV 16 plasmid to samples of human placental DNA at 1 ng/μL and reported an amplification of 26-fold for human DNA and 679-fold for HPV 16 DNA, suggesting that MDA is a good method for enriching circular HPV genomes [5]. Using MDA and NGS sequencing, researchers have been able to detect a plethora of novel HPVs as well as known HPV types, not detected by traditional PCR-based enrichment methods, among skin lesions/tumors and condyloma accuminata [3,4,7].

Another enrichment approach is to use a set of specific probes, "baits", to recover HPV sequences from the entire genome of the virus. In brief, labelled biotinylated HPV specific probes are captured by streptavidin coated magnetic beads after hybridization, resulting in "pure" HPV-derived reads. Consistent with the overlapping PCR approach, these probes need to be designed in advance but can be modified as required if low read numbers are obtained for some regions of the viral genome. Although it can be expensive,

the advantage of the probes/baits approach is the large number of different probes one may include, allowing thousands of probes in one design. This means that it is possible to load the analysis with probes for all known HPVs, depending on purpose. Furthermore, it is the current gold standard method for integration studies [52].

Regardless of protocol applied, to validate the quality of host depletion or viral enrichment, positive and negative control material must be added at the nucleic acid extraction step and carried through the enrichment/capture/depletion step and all stages of subsequent analysis. As a positive control, cell line material infected with HPV is commonly used. However, we would recommend using specimens that contain both human DNA and HPV DNA in the typical concentrations that would correspond to real-life clinical specimens (99:1), as the sensitivity of the different approaches may vary. As a negative control, human DNA (HPV free, but containing the corresponding background "noise" as a true HPV-negative sample) or DNA-free water can be used. Note that if primer-based target enrichment is used, negative controls may contain noise such as primer-dimer bands when assessed by electrophoreses as well as unspecific amplification; these should be clearly discriminated from target sequences. Again, the introduction of homogeneous internal quality controls (IQCs) in the nucleic acid amplification should help in the identification of such issues. Internal positive control material demonstrates whether the depletion/lysis has removed/reduced the HPV target in large proportion during the sample preparation stage. For the enrichment step, positive control material should reflect if amplification of the target regions/genome occurs and negative controls help determine whether any nonexpected amplification/targeting occurred.

#### 2.1.4. Direct Sequencing

Several operators have opted for performing WGS directly after nucleic acid extraction, without enrichment. As described above, reaching a high sequencing depth is required, due to the low proportion of viral sequences typically present in the human specimens. Direct sequencing has enabled an agnostic approach to DNA presence in clinical samples: while around 10% of cervical cancers are found to be negative for oncogenic HPVs by traditional PCR-based genotyping methods [57–61]; direct sequencing can reveal the presence of viral sequences of potentially causative HPVs in such cases [13,62]. Interestingly, most of the HPVs detected with direct sequencing among the carcinomas negative by traditional PCR typing systems corresponded to HPV types within the explicit detection range of these assays. Therefore, the information obtained by direct sequencing can be used, not only to detect novel variants or types that may have escaped traditional amplification but also as a way to quantify and monitor shortfalls in detection due to sensitivity issues. Furthermore, evidence suggests that HPV-negative cancer patients have a worse longitudinal prognosis [60,63,64]; this is reflected in the staging system for oropharyngeal cancer which acknowledges the dichotomous disease status based on HPV presence [65] as does the recent WHO update for female genital tumors [66]. How "best" to annotate HPV status in cancer tissue is an area which arguably lacks consensus in the literature; however, NGS provides a powerful tool to at least resolve which cancer cases may be truly virally negative.

#### 2.1.5. Library Preparation and Sequencing

At present, there are several different sequencing chemistries available. Each system has its own protocols and due to the diversity of platforms, and rapid pace of developments, we cannot recommend a specific one for all HPV applications. Akin to the assessment and introduction of any new technology in-house, we strongly recommend validation that includes confirmation of expected results from "known" quality materials and the evaluation of different kits with the specific analytical purpose in mind. For laboratories looking to embed NGS into the accredited scope of their clinical service, initial validation followed by yearly verification would likely be mandatory.

Quality, quantity, and fragment length analysis is a must to confirm success of library preparation. Library preparation protocols usually inform the operator about the concentration and size expected for prepared libraries. If measurements do not reach the expected values, (e.g., fragments are too big/small or the library concentration is too diluted) optimization of fragmentation times, clean-up processes, or amplification steps should be performed. Larger fragments cluster less efficiently than smaller molecules, and a low concentration of prepared libraries translates into a low number of sequenced reads. Here, it is crucial to consider in which context libraries are analyzed; in a research study, suboptimal measurements can occasionally be acceptable but probably never in a clinical context where protocol adherence is paramount and actionable test results are needed.

One also needs to perform accurate normalization to obtain homogeneous distribution (number of sequencing reads) of the samples and assure that the proper sequencing read length is used, depending on the insert size of the library. As an example, Illumina libraries prepared with dual-indexing that show a fragment length of 200 bp should not be sequenced with 2 × 150 bp, as part of the fragment length (around 130 bp) corresponds to adapter sequences, and the insert size, which is the actual query sequence, only comprises 70 bp (200 − 130 bp). Thus, 80 bp of the 150 bp sequenced (150 − 70 bp) does not contribute useful information. Libraries from positive and negative samples must be added into the final input dilution. Our recommendations for quality-control steps in the NGS workflow are summarized in Table 1.

**Table 1.** Steps, potential quality issues, and proposed mitigations for next-generation sequencing (NGS) analytical workflow, with a focus on whole-genome sequencing purposes.

| Human Papillomavirus Detection by Next-Generation Sequencing | | |
|---|---|---|
| **NGS Step** | **Possible Difficulties** | **Mitigations** |
| Sample preparation | Nucleic acid quality and/or quantity outside library prep kit requirements | Selection of appropriate nucleic acid extraction methods.<br>Pilot study comparing different extraction kits.<br>Selection of enrichment or depletion protocol.<br>Introduction of homogeneous internal quality controls |
| | Incorrect fragment size (too short or too long) | Electrophoresis, bioanalyzer, and/or fluorometric quantitation.<br>Selection of further protocols based on the fragment size (e.g., use of shorter amplicons if DNA is highly fragmented) |
| Library preparation and sequencing | Incorrect fragment size (too short or too long)<br>Incorrect number of sequencing reads or partial reads | Correct selection of library kit and fragment length.<br>Correct selection of sequencing kit (e.g., 75 bp and 150 bp) to avoid sequencing adapters or longer fragments that insert size. |
| Data analysis | Low sequencing depth | Library preparation and sequencing piloting, and re-analysis<br>Confirm reference sequence is correct.<br>Use of updated database.<br>In case of low sequencing depth at the beginning or end of the reference sequence, note that HPV is circular and not linear as the reference. |
| | Incorrect alignment | Confirmation that desired alignment cut-offs are correct. |
| | Mix/chimeras of microbial organisms | Filter reads—use of updated databases and careful settings of parameters. De novo assembly evaluation (HPV Chimera scripts) |
| | Validation of pipeline | Digital IQC, EQA, external assessment.<br>Interlaboratory comparison |
| Storage | Large amount of data | Cloud services, compression of files, and storage of only raw input and final output. |
| | Security | Restricted super-user access, individually curated data access<br>Analyst working only with coded/pseudonymized samples (where an independent database administrator holds the key code at another site) |
| | Length of data storage | Organization policy/data archiving laws and regulations |

## *2.2. NGS-Process Step 2: Bioinformatical Analysis*

### 2.2.1. Raw Sequence Data Management

The output data from a sequencing machine are often referred to as raw data. Raw data management generally includes filtering steps to remove poor-quality data and host-derived human reads and continues by mapping nonhuman high-quality reads directly to a

known reference database or performing a de novo assembly approach, finishing with HPV taxonomy classification, phylogenetic analysis, and variant calling. There are several open-access tools that can be used to analyze "big data". A set of bioinformatic algorithms, when executed in a predefined sequence, is collectively referred to as a bioinformatics "pipeline". These pipelines can be designed in-house by teams with available bioinformatics expertise or obtained as ready-to-use applications from commercial suppliers. The premade pipelines are principally aimed to users with little bioinformatic experience and act as a "blackbox" (user does not know which algorithm and calculation(s) are used by the pipeline).

#### 2.2.2. Sequence Analyses: Quality Assessment of Reads

Each bioinformatical tool deployed in the pipeline uses different algorithms and parameters when handling data. When the objective of a project is to resolve a 0.5–1% difference between sequences, these differences in parameter settings could mean a different interpretation on nucleotide/mutation level is reached depending on which pipeline is being used. Ergo, two studies on the same nucleotide position could reach two different conclusions as to whether a mutation is present.

Errors due to poor sample handling and storage conditions, polymerase bias, PCR- or qPCR-induced errors, and incorporation errors within sequencing may be introduced during sample preparation, amplification, library preparation, and sequencing stages. While these errors might not interfere with the identification of an HPV type (where the sequence divergence is 10% relative to its most closely related type), they might compromise the identification of sublineages or variations in nucleotides that could have implications on accuracy, consistency, and, potentially, predicted phenotype. Careful sample handling, selection of a high-fidelity polymerase [67], and quality control of the raw data is a must [68].

Most analysis applications use FASTQ files as input for analysis; however, different sequencing instruments may give different extensions for raw sequencing data (e.g., Illumina generates bcl files), and the first step is the conversion of those files to a standard format (FASTQ). Platforms usually provide software for the desired conversion (e.g., bcl2fastq from Illumina). Raw FASTQ files should be subjected to quality trimming and adaptor removal as a first step. Software for quality trimming and adaptor removal include Cutadapt, Trimmomatic, Trim Galore!, SeqTrim, and FastX among others [69]. Quality trimming is performed to remove low quality reads and aims to reduce the effect of the progressive decrease in sequencing quality with the increased length of the sequenced library. Trimming removes low quality portions of NGS reads while preserving the high-quality part of such a read. The user can specify the quality cut-off for a base or use a "sliding window" approach (defined as setting a cut-off for the average quality detected in a number of X contiguous bases instead of just one base). Quality is usually checked according to the Phred quality scores, which are scores logarithmically related to base-calling error probabilities [70]. As an example, a Phred quality score of Q30 corresponds to a base calling accuracy of 99.9% (1 error per 1000 bp). The minimum quality recommended in the literature is a Phred quality score of 20 (99.0% accuracy; 1 error per 100 bp), with the optimal quality however being above Q30 [71].

#### 2.2.3. Alignment of Reads to a Suitable Human Reference Genome

To obtain a dataset that contains only reads of interest, e.g., viral-related reads for HPV detection, nontarget sequences may be filtered out at the bioinformatics level to speed up downstream analysis and decrease the risk of misassemblies of genomic data. Most researchers applying WGS from sequenced extracted material opt for filtering out human genome sequences only (leaving HPV plus "other" microorganism reads). This is practical as human reads (according to our experience of multiple sequencing projects with a metagenomic perspective) account for approximately 90% of the total (with some variability depending on the sample origin) [7,13,62]. While several successive "versions" corresponding to human genome reference exist, it is recommended that the latest build

(released in December of 2013), officially named GRCh38 (Genome Research Consortium human build 38) or commonly Hg38 (human genome build 38) is used.

Numerous aligners exist so far and are being developed in order to achieve greater accuracy pertaining to precision. Widely used tools include BWA-MEM (Burrows-Wheeler aligner) [72], SOAP2 [73], or Bowtie 2 [74], but several commercial tools are also available such as NextGenmap and Novoalign [75,76].

While some operators may, at this point, choose to filter out reads that are identical to the human reference genome (100% identical), another approach is to employ looser parameters which accept reads as human if the identity and coverage across the human genome sequence of interest are at least 95% and 75%, respectively. This latter approach allows for the detection of possible mutations not anticipated in the reference sequence, although this flexibility should be tempered so as not to misclassify nonhuman reads as human. An important thing to consider after performing the aligning/mapping to the human reference genome is to select which "unmapped" reads are to be used in downstream analysis. Sequencing with paired-end reads may contain (1) paired-end reads where both reads are unmapped, (2) paired-end reads where one of the pair read maps to the genome and the other does not. Operators may decide to discard nonhuman single reads and continue only with nonhuman paired-end reads, or to include them all; such a choice is entirely dependent on coverage obtained and the aim of the project/resolution required.

#### 2.2.4. Alignment of Reads to a Suitable HPV Reference Database

Once human (or host) sequences are filtered out from the high-quality data set, most operators align reads to a known and curated HPV database for HPV classification or to the reference HPV genome in question. Currently, there are 222 different HPV types officially established (data accessed on 18 March 2021 from Hpvcenter.se) and another 220 putative novel HPV types (not cloned and investigated by the International HPV Reference Center) whose complete sequence can be found in the public database from the papillomavirus Episteme (data accessed on 18 March 2021 at https://pave.niaid.nih.gov/).

Furthermore, there are many partial genomic sequences of HPV isolates (not all specifying which HPV type they correspond to) available at public databases, with GenBank having >33,500 hits retrieved when typing "human papillomavirus" (data accessed on 18 March 2021). A recent publication [77] detected up to 0.5% chimeric sequences and/or taxonomy errors when analyzing HPV sequences in the GenBank database. This highlights the importance that the database be obtained from a quality reference repository and that local curation is performed before doing any type of alignment, such as checking for potential errors and updates. In-house databases belonging to individual investigators should be periodically updated with canonical or reference types, as new HPV types are continuously being discovered [5,78].

It is particularly key to note that, when aligning the sequencing reads to a specific HPV genotype, operators ascertain that the correct reference sequence is used. There are 3650 sequences belonging to HPV16 isolates in GenBank (sequence length 7500–8500, data accessed on 18 March 2021), showing differences that may reach up to 10% of the total genome. If each separate investigation were to use a different sequence as reference genome, then comparison between publications becomes challenging at best. The reference genomes that should be used for each HPV type are provided at the International HPV Reference Laboratory website (Hpvcenter.se); accessed 7 July 2021, as well as at the papillomavirus Episteme database (https://pave.niaid.nih.gov/; accessed 7 July 2021). The latter resource has a contemporary collection of internationally ratified sequences from the reference clones corrected for known sequencing mistakes in the original sequences.

#### 2.2.5. Identification of HPV Types/Lineage/Sublineages

Classification of HPVs is based on the nucleotide sequence homology of the *L1* gene, which is the most conserved region of the viral genome. Within the family, different

genera share less than 60% nucleotide similarity. Within each genus, different species share between 60% and 70% similarity. Below the species level, a novel HPV type shares less than 90% similarity to any other type [79–81]. The definition of a variant lineage is that the L1 open-reading frame differs by more than 1%, but less than the 10% that would make it another HPV type [82]. A variant sublineage is defined as groups of sequences with 0.5–1.0% differences between genomes [83].

There exist different tools for the identification of variants. One of the most used and user friendly is BLAST [84]. This tool compares the sequence under investiagation to sequences stored in the database, detailing statistical significance of matches. Again, the importance of using standard references for HPV variants is essential. Burk et al. described the representative genomes for viral variant lineages and sublineages, and most authors rely on these sequences as variant lineages references [82].

If a phylogenetic analysis is required, different open-source tools exist (RaxML, MegaX) that infer phylogenetic trees after choosing the statistical method [85,86].

#### 2.2.6. Evaluation of Coverage across the Genome

If the purpose is to detect HPV genotypes (not within-genotype specific variant calling), once the reads are aligned to the HPV database, we recommended that cut-offs are applied on which HPV positivity is based (Table 1).

This could be, e.g., setting a minimum of 10 reads detected for a specific HPV type together with a coverage of at least 10% of the HPV genome (around 800 bp coverage). This approach would avoid false positivity generated by background noise (e.g., presence of many low complexity reads mapping to just a small region of the genome). If phylogenetic analysis or variant calling is required, a FASTA file with the "query" sequence must be created. When creating a FASTA file from the obtained sequencing reads/contigs, investigators should be aware of the extent of genome coverage to see if there are missing regions

Evaluating the full coverage of the sequence is important, as several tools that convert the sequencing reads into a FASTA file use a reference sequence to account for the regions that are not covered. The use of "N"s is recommended for the positions that are not covered by the sequencing reads. For variant lineage assignment, exclusion must be considered for specimens with poor read depth (<200 median depth) and/or low genome coverage (<80% genome coverage) [53]. For variant calling, even stricter cutoffs should be applied. Additionally, further steps including marking duplicates to identify read pairs likely to have originated from duplicates of the same original DNA fragments and recalibration of base quality scores should be performed, as suggested by the best practices at GATK [87]. Considering just the base depth as a cutoff for variant calling (e.g., five reads per position) is not enough to assure accurate calling. It is essential to differentiate between true positive variants and false positive variants. Parameters and statistics which describe how many reads cover the variant, what proportion of reads are in forward vs. reverse orientation, and what the sequence context is like around the variant site should be considered.

#### 2.2.7. De Novo Assembly of HPV Contigs

When the correct reference genome is not known, the (re-)construction of the sequenced genome must be performed without a priori knowledge of either the correct original sequence (or the order of the DNA fragments), by assembling overlapping reads into one or more contigs. This process is known as de novo assembly. Subsequent post assembly assessment is mandatory to reduce the risk of chimeric sequences and possible miscalling of HPV positivity in samples and/or erroneous calling of new HPV variants/genotypes. HPV-Chimera scripts exist to help researchers determine the accuracy of their HPV contigs [12,77].

#### 2.2.8. Digital Quality Assessment

While positive and negative controls can be incorporated into laboratory experiments and several quality check-points are available during the whole laboratory process, we cur-

rently lack an agreed approach for the quality control of bioinformatical tools and pipelines. Digital IQCS can be prepared from confirmed and verified positive material and stored as FASTQ files (raw data). Interlaboratory exchange of data can provide reassurance by comparing results on sequences derived from the same specimens. However, this requires resources and collaboration which may not always be available in the short term. Therefore, it would be beneficial to have positive IQCS available in an online repository that could be used to verify the pipeline when setting up a new service/test or when a tool is updated.

#### 2.2.9. Journal Submission Requirements

Recently, several journals have started to request that all authors who submit manuscripts containing NGS data provide a detailed summary of sequencing coverage and quality statistics. For example, the International Journal of Cancer requires a summary from submitting authors that must include all information about library preparation, sequencing technology information (e.g., platform, read length, and paired-end/single read approach), as well as preprocessing, quality control, and filtering of the raw NGS data [88]. Furthermore, the sequencing coverage and quality statistics of each sample must be summarized as a Supplementary Table.

#### 2.2.10. HPV NGS in Clinical Settings

Application of NGS for clinical testing requires a level of quality assurance and monitoring likely to be even more stringent than systems set up in research laboratories. Verification and validation for each of the steps that make up the NGS process is the key to obtain and provide reliable results to clinicians and patients. Any minor change of the wet-lab protocol or any parameter in the data analysis requires a full verification with previously known samples/sequences. This in combination with the external quality assessment, and accreditation helps ensure the validity of the clinical results.

#### 2.2.11. External Quality Assessment and Accreditation

Suppliers of EQA schemes have developed external quality materials to support NGS sequencing results for various pathogens and human genes (e.g., GenQA, Statens Serum Institut, and QCMD (pilot)). At time of publication, we are not aware of any official HPV EQA scheme to support NGS for wet and in silico analysis or indeed data/dry analysis. This is arguably a current deficit, as while interlaboratory exchange of materials is undoubtedly helpful for quality assurance and validation of HPV NGS, such exchanges do not wholly "stand in" for consistent performance in a formal accredited EQA scheme(s). Should HPV NGS move to the diagnostic context, then this would increasingly require address.

#### 2.2.12. Data Storage Requirements

Data storage demands of NGS are often very large and need to be carefully considered before local implementation. Unfortunately, there is no (international) consensus on what and how data should be stored beyond the general recommendation that it should be stored in line with national and local capacity and governance policies. In high-level terms, we recommend storing the raw FASTQ files in a compressed mode (.fastq.gz), the final output, and the full-log file (documentation on the software/tools used, including versions, parameters, and github location, needed to obtain the output files) making the whole analysis reproducible.

If intermediate files are to be kept, they should be stored as standard open-file formats FASTQ, BAM, and VCF, facilitating the exchange with other laboratories, where governance permits. Cloud-based storage could be a very helpful tool; however, this may be challenging to reconcile with data protection. Currently, most journals ask for data availability and request that authors upload all nonhuman sequences detected in the study to different databases (such as the European Nucleotide Archive (ENA), Sequencing Read Archive (SRA), and Genbank) to make research publicly available and re-usable for other scientists without compromising confidentiality.

## 3. Discussion

Next-generation sequencing (NGS) has enabled researchers to detect human papillomavirus (HPV) infections with unprecedented sensitivity and accuracy while simultaneously providing the whole viral sequence to be analyzed for viral point mutations, variant lineages, and genome variations. Increasing incorporation of NGS into laboratories for research, epidemiology, and diagnostic purposes may be only a matter of time, particularly as costs reduce with increasing demand and competition.

Certainly, the use of NGS for clinical workstreams generally requires accreditation/auditing from an independent regulator. As an example, the National Accreditation Body for the United Kingdom (UKAS) now has expertise to assess and accredit labs that have NGS in scope, working to ISO 15189:2012 standards. This process covers assessment of staff training, quality control of pipelines, initial validation of the whole process compared to a gold-standard approach, and reproducibility of results plus checks to ensure security of data and access is correct. The magnitude of the effort to achieve accreditation in a particular service laboratory is likely to depend on local support for the integration of NGS to support public health epidemiology and precision medicine in general terms. The SARS-CoV-2 pandemic has brought about an exponential increase in molecular testing and associated infrastructure, including that required to support sequence-level variant detection. This is likely to pay dividends for the establishment of "cross organism working" to a diagnostic standard [89]. Certainly, NGS has the potential to support risk stratification of patients with HPV-associated disease through its ability to detect types within tissue with exquisite sensitivity and through providing subtle insights into aspects of HPV infection that may be predictive of outcome (integration pattern and status, delineation of dominant variant(s), and viral load). NGS may also be applied to liquid biopsies, from blood, to support longitudinal monitoring of treatment success [90].

The bulk of HPV sequencing to date has been performed using the IonTorrent initially [9,11,91], followed by the newer Illumina platform [7,53,57]; some very recent studies have also used nanopore platforms [52,92]. While many studies report higher sensitivity and accuracy when comparing NGS to routine genotyping methods [6,8], there are very few studies where one HPV NGS approach has been directly compared to another in a head-to-head approach using the same sample set [93,94]. Hence, there are fewer data available compared to the relative wealth of peer-reviewed literature on head-to-head performance of traditional HPV molecular assays. While many researchers are already using NGS for detection and analysis of HPV-associated diseases, there is no clear international consensus about what steps are to be performed nor which quality criteria are appropriate.

Typically, published presequencing laboratory protocols appear well validated for providing detail on clear-quality assessment procedures and specifying the requirements needed for library preparation evaluation and success. Nevertheless, even though the inclusion of positive and negative controls should be a must, the presence and/or description of these controls is not commonly found within NGS publications. This may be in part due to the relatively high cost of NGS. Hornung et al. reviewed ~265 publications which had used NGS for microbiome research and found that only 30% of publications used any type of negative controls and that less than 10% reported positive controls. Additionally, they observed that some of the results reported were potentially indistinguishable from contaminants [95].

Furthermore, to the best of our knowledge, quality assessment of bioinformatic analysis appears not to be as standardized, when compared to the analytical sample handling stages. However, we consider it key that the use of a curated and updated databases, use of standard reference genome sequences, and description of the parameters used for each step should be described fully in all publications.

The present piece aims to describe and summarize the application of WGS for HPV detection and has mainly focused on DNA detection to detect the whole HPV sequence. Nevertheless, RNA sequencing should not be forgotten as an alternative method for HPV detection, as RNA transcription not only enables detection of HPVs but provides further

information on the active HPV infection that drives viral oncogene expression. In the cancers that are known to be caused by HPV, transcription of viral genes is necessary for viral pathogenicity [96]. Studying transcription of the E6 and E7 oncogenes has been useful to elucidate which infections are likely to be involved in the etiology of the tumor, e.g., in head and neck cancer studies [97–99]. Even though RNA sequencing is not usually performed within studies aiming to analyze the whole HPV genome (as noncoding regions, e.g., URR, are not detected if RNA extraction and sequencing are performed), important information about active infection and viral oncogene expression is obtained with this sapproach.

International collaboration is essential to efficiently further knowledge, scientific development, and concerted efforts to combat globally prevalent viral infections. In the case of HPV, the International HPV Laboratory Network LabNet was created by WHO in 2006 to support global development of laboratory standardization and quality assurance of HPV detection methods. LabNet concentrated on evaluating and improving methods used for research and evaluation of HPV-based screening and vaccination [100]. As part of this effort, LabNet published an HPV laboratory manual, based on knowledge and experience gained through international collaborative studies, aiming to assist in establishing the laboratory support required for HPV research [101]. The successor to LabNet, the International HPV Reference Center (Hpvcenter.se), aims to support reliable and comparable HPV detection services, allowing data to be internationally comparable. The Center has organized and issued global proficiency panels (sets of blinded samples containing HPV genotypes at different concentrations) since 2008, and a definite improvement in average assay performance globally has been seen since the panel was issued [102]. The SHPVRL has also acted as a hub laboratory to support the creation of materials and best practice documents to facilitate the introduction of new HPV technologies and their continued monitoring [103,104].

Together. we now work between our two laboratories to exchange samples, know-how, and protocols for bioinformatical flow and sequence analyses. Hopefully, this will strengthen the quality of work produced by both settings and act as a catalyst/model for future international endeavors. We propose that it is time for NGS to be included in the global efforts on quality assurance and improvement in HPV-based testing and diagnostics. By establishing a set of quality standards and best-practice statements, the community could systematically develop and apply NGS guidelines suited for HPV research, epidemiology, and diagnostics to ensure this innovative and powerful technology is developed in an internationally comparable and robust manner.

**Author Contributions:** Conceptualization, K.C. and K.S.; methodology, L.S.A.M. and D.G.; writing—original draft preparation, L.S.A.M. and D.G.; writing—Review and Editing, K.C. and K.S.; supervision, K.C. and K.S.; funding acquisition, L.S.A.M. and K.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the Center for Innovative Medicine (CIMED grant number 613/06, to KS), the Swedish Medical Society (SLS, grant number 885941, to KS), and the Swedish Foundation for Strategic Research (grant number RB13-0011, supporting KS and SAM). The SHPVRL is supported by National Services Division of the National Health Service in Scotland.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest related to this work.

## References

1. Lorincz, A.; Wheeler, C.M.; Cuschieri, K.; Geraets, D.; Meijer, C.J.L.M.; Quint, W. Developing and Standardizing Human Papillomavirus Tests. In *Human Papillomavirus: Proving and Using a Viral Cause for Cancer*; David Jenkins, F.X.B., Ed.; Academic Press: Cambridge, MA, USA, 2020; pp. 111–130.

2. Gradissimo, A.; Burk, R.D. Molecular tests potentially improving HPV screening and genotyping for cervical cancer prevention. *Expert Rev. Mol. Diagn.* **2017**, *17*, 379–391. [CrossRef]
3. Arroyo Muhr, L.S.; Bzhalava, D.; Lagheden, C.; Eklund, C.; Johansson, H.; Forslund, O.; Dillner, J.; Hultin, E. Does human papillomavirus-negative condylomata exist? *Virology* **2015**, *485*, 283–288. [CrossRef]
4. Arroyo Muhr, L.S.; Hultin, E.; Bzhalava, D.; Eklund, C.; Lagheden, C.; Ekstrom, J.; Johansson, H.; Forslund, O.; Dillner, J. Human papillomavirus type 197 is commonly present in skin tumors. *Int. J. Cancer* **2015**, *136*, 2546–2555. [CrossRef]
5. Bzhalava, D.; Muhr, L.S.; Lagheden, C.; Ekstrom, J.; Forslund, O.; Dillner, J.; Hultin, E. Deep sequencing extends the diversity of human papillomaviruses in human skin. *Sci. Rep.* **2014**, *4*, 5807. [CrossRef]
6. Arroyo, L.S.; Smelov, V.; Bzhalava, D.; Eklund, C.; Hultin, E.; Dillner, J. Next generation sequencing for human papillomavirus genotyping. *J. Clin. Virol.* **2013**, *58*, 437–442. [CrossRef] [PubMed]
7. Bzhalava, D.; Johansson, H.; Ekstrom, J.; Faust, H.; Moller, B.; Eklund, C.; Nordin, P.; Stenquist, B.; Paoli, J.; Persson, B.; et al. Unbiased approach for virus detection in skin lesions. *PLoS ONE* **2013**, *8*, e65953. [CrossRef]
8. Nilyanimit, P.; Chansaenroj, J.; Poomipak, W.; Praianantathavorn, K.; Payungporn, S.; Poovorawan, Y. Comparison of Four Human Papillomavirus Genotyping Methods: Next-generation Sequencing, INNO-LiPA, Electrochemical DNA Chip, and Nested-PCR. *Ann. Lab. Med.* **2018**, *38*, 139–146. [CrossRef] [PubMed]
9. Cullen, M.; Boland, J.F.; Schiffman, M.; Zhang, X.; Wentzensen, N.; Yang, Q.; Chen, Z.; Yu, K.; Mitchell, J.; Roberson, D.; et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res.* **2015**, *1*, 3–11. [CrossRef] [PubMed]
10. Clifford, G.M.; Tenet, V.; Georges, D.; Alemany, L.; Pavon, M.A.; Chen, Z.; Yeager, M.; Cullen, M.; Boland, J.F.; Bass, S.; et al. Human papillomavirus 16 sub-lineage dispersal and cervical cancer risk worldwide: Whole viral genome sequences from 7116 HPV16-positive women. *Papillomavirus Res.* **2019**, *7*, 67–74. [CrossRef]
11. Mirabello, L.; Yeager, M.; Cullen, M.; Boland, J.F.; Chen, Z.; Wentzensen, N.; Zhang, X.; Yu, K.; Yang, Q.; Mitchell, J.; et al. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J. Natl. Cancer Inst.* **2016**, *108*. [CrossRef] [PubMed]
12. Arroyo Muhr, L.S.; Lagheden, C.; Hassan, S.S.; Kleppe, S.N.; Hultin, E.; Dillner, J. De novo sequence assembly requires bioinformatic checking of chimeric sequences. *PLoS ONE* **2020**, *15*, e0237455. [CrossRef]
13. Arroyo Muhr, L.S.; Lagheden, C.; Lei, J.; Eklund, C.; Nordqvist Kleppe, S.; Sparen, P.; Sundstrom, K.; Dillner, J. Deep sequencing detects human papillomavirus (HPV) in cervical cancers negative for HPV by PCR. *Br. J. Cancer* **2020**, *123*, 1790–1795. [CrossRef]
14. Perez, S.; Cid, A.; Araujo, A.; Lamas, M.J.; Saran, M.T.; Alvarez, M.J.; Lopez-Miragaya, I.; Gonzalez, S.; Torres, J.; Melon, S. A novel real-time genotyping assay for detection of the E6-350G HPV 16 variant. *J. Virol. Methods* **2011**, *173*, 357–363. [CrossRef] [PubMed]
15. Mirabello, L.; Yeager, M.; Yu, K.; Clifford, G.M.; Xiao, Y.; Zhu, B.; Cullen, M.; Boland, J.F.; Wentzensen, N.; Nelson, C.W.; et al. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* **2017**, *170*, 1164–1174.e6. [CrossRef]
16. Lee, J.Y.; Cutts, R.J.; White, I.; Augustin, Y.; Garcia-Murillas, I.; Fenwick, K.; Matthews, N.; Turner, N.C.; Harrington, K.; Gilbert, D.C.; et al. Next Generation Sequencing Assay for Detection of Circulating HPV DNA (cHPV-DNA) in Patients Undergoing Radical (Chemo)Radiotherapy in Anal Squamous Cell Carcinoma (ASCC). *Front. Oncol.* **2020**, *10*, 505. [CrossRef]
17. Besser, J.; Carleton, H.A.; Gerner-Smidt, P.; Lindsey, R.L.; Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.* **2018**, *24*, 335–341. [CrossRef]
18. Gargis, A.S.; Kalman, L.; Lubin, I.M. Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. *J. Clin. Microbiol.* **2016**, *54*, 2857–2865. [CrossRef] [PubMed]
19. Lopez-Labrador, F.X.; Brown, J.R.; Fischer, N.; Harvala, H.; Van Boheemen, S.; Cinek, O.; Sayiner, A.; Madsen, T.V.; Auvinen, E.; Kufner, V.; et al. Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. *J. Clin. Virol.* **2021**, *134*, 104691. [CrossRef] [PubMed]
20. Endrullat, C.; Glokler, J.; Franke, P.; Frohme, M. Standardization and quality management in next-generation sequencing. *Appl. Transl. Genom.* **2016**, *10*, 2–9. [CrossRef]
21. Scottish Science Advisory Council. Informing the Future of Genomic Medicine in Scotland. Available online: https://www.scottishscience.org.uk/sites/default/files/article-attachments/Genomics%20Full%20Report.pdf (accessed on 26 May 2021).
22. Medlineplus. Available online: Medlineplus.gov/genetics/understanding/precisionmedicine/definition/ (accessed on 26 May 2021).
23. Wong, S.Q.; Li, J.; Tan, A.Y.; Vedururu, R.; Pang, J.M.; Do, H.; Ellul, J.; Doig, K.; Bell, A.; MacArthur, G.A.; et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC Med. Genom.* **2014**, *7*, 23. [CrossRef]
24. Yost, S.E.; Smith, E.N.; Schwab, R.B.; Bao, L.; Jung, H.; Wang, X.; Voest, E.; Pierce, J.P.; Messer, K.; Parker, B.A.; et al. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **2012**, *40*, e107. [CrossRef] [PubMed]
25. Kerick, M.; Isau, M.; Timmermann, B.; Sultmann, H.; Herwig, R.; Krobitsch, S.; Schaefer, G.; Verdorfer, I.; Bartsch, G.; Klocker, H.; et al. Targeted high throughput sequencing in clinical cancer settings: Formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genom.* **2011**, *4*, 68. [CrossRef]

26. Graw, S.; Meier, R.; Minn, K.; Bloomer, C.; Godwin, A.K.; Fridley, B.; Vlad, A.; Beyerlein, P.; Chien, J. Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci. Rep.* **2015**, *5*, 12335. [CrossRef] [PubMed]
27. Nanodrop. Technical Support Bulletin. Available online: https://bio.davidson.edu/projects/gcat/protocols/NanoDrop_tip.pdf (accessed on 26 May 2021).
28. Illumina. Nextera®DNA Library Prep Reference Guide. Available online: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nexteradna/nextera-dna-library-prep-reference-guide-15027987-01.pdf (accessed on 26 May 2021).
29. Do, H.; Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clin. Chem.* **2015**, *61*, 64–71. [CrossRef]
30. Bettoni, F.; Koyama, F.C.; De Avelar Carpinetti, P.; Galante, P.A.F.; Camargo, A.A.; Asprino, P.F. A straightforward assay to evaluate DNA integrity and optimize next-generation sequencing for clinical diagnosis in oncology. *Exp. Mol. Pathol.* **2017**, *103*, 294–299. [CrossRef]
31. Duncavage, E.J.; Magrini, V.; Becker, N.; Armstrong, J.R.; Demeter, R.T.; Wylie, T.; Abel, H.J.; Pfeifer, J.D. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J. Mol. Diagn.* **2011**, *13*, 325–333. [CrossRef]
32. Allander, T.; Emerson, S.U.; Engle, R.E.; Purcell, R.H.; Bukh, J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 11609–11614. [CrossRef]
33. Duhaime, M.B.; Sullivan, M.B. Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* **2012**, *434*, 181–186. [CrossRef]
34. Depledge, D.P.; Palser, A.L.; Watson, S.J.; Lai, I.Y.; Gray, E.R.; Grant, P.; Kanda, R.K.; Leproust, E.; Kellam, P.; Breuer, J. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE* **2011**, *6*, e27805. [CrossRef] [PubMed]
35. Koehler, J.W.; Hall, A.T.; Rolfe, P.A.; Honko, A.N.; Palacios, G.F.; Fair, J.N.; Muyembe, J.J.; Mulembekani, P.; Schoepp, R.J.; Adesokan, A.; et al. Development and evaluation of a panel of filovirus sequence capture probes for pathogen detection by next-generation sequencing. *PLoS ONE* **2014**, *9*, e107007. [CrossRef]
36. Wylie, T.N.; Wylie, K.M.; Herter, B.N.; Storch, G.A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **2015**, *25*, 1910–1920. [CrossRef]
37. Ji, X.C.; Zhou, L.F.; Li, C.Y.; Shi, Y.J.; Wu, M.L.; Zhang, Y.; Fei, X.F.; Zhao, G. Reduction of Human DNA Contamination in Clinical Cerebrospinal Fluid Specimens Improves the Sensitivity of Metagenomic Next-Generation Sequencing. *J. Mol. Neurosci.* **2020**, *70*, 659–666. [CrossRef] [PubMed]
38. Hasan, M.R.; Rawat, A.; Tang, P.; Jithesh, P.V.; Thomas, E.; Tan, R.; Tilley, P. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. *J. Clin. Microbiol.* **2016**, *54*, 919–927. [CrossRef]
39. Gao, G.; Wang, J.; Kasperbauer, J.L.; Tombers, N.M.; Teng, F.; Gou, H.; Zhao, Y.; Bao, Z.; Smith, D.I. Whole genome sequencing reveals complexity in both HPV sequences present and HPV integrations in HPV-positive oropharyngeal squamous cell carcinomas. *BMC Cancer* **2019**, *19*, 352. [CrossRef] [PubMed]
40. Wentzensen, N.; Vinokurova, S.; Von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.* **2004**, *64*, 3878–3884. [CrossRef]
41. Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. *Nature* **2017**, *543*, 378–384. [CrossRef] [PubMed]
42. Chandrani, P.; Kulkarni, V.; Iyer, P.; Upadhyay, P.; Chaubal, R.; Das, P.; Mulherkar, R.; Singh, R.; Dutt, A. NGS-based approach to determine the presence of HPV and their sites of integration in human cancer genome. *Br. J. Cancer* **2015**, *112*, 1958–1965. [CrossRef] [PubMed]
43. Zhang, R.; Shen, C.; Zhao, L.; Wang, J.; McCrae, M.; Chen, X.; Lu, F. Dysregulation of host cellular genes targeted by human papillomavirus (HPV) integration contributes to HPV-related cervical carcinogenesis. *Int. J. Cancer* **2016**, *138*, 1163–1174. [CrossRef] [PubMed]
44. Ibragimova, M.; Tsyganov, M.; Shpileva, O.; Churuksaeva, O.; Bychkov, V.; Kolomiets, L.; Litviakov, N. HPV status and its genomic integration affect survival of patients with cervical cancer. *Neoplasma* **2018**, *65*, 441–448. [CrossRef] [PubMed]
45. Shen, C.; Liu, Y.; Shi, S.; Zhang, R.; Zhang, T.; Xu, Q.; Zhu, P.; Chen, X.; Lu, F. Long-distance interaction of the integrated HPV fragment with MYC gene and 8q24.22 region upregulating the allele-specific MYC expression in HeLa cells. *Int. J. Cancer* **2017**, *141*, 540–548. [CrossRef]
46. Koneva, L.A.; Zhang, Y.; Virani, S.; Hall, P.B.; McHugh, J.B.; Chepeha, D.B.; Wolf, G.T.; Carey, T.E.; Rozek, L.S.; Sartor, M.A. HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers. *Mol. Cancer Res.* **2018**, *16*, 90–102. [CrossRef]
47. Han, L.; Maimaitiming, T.; Husaiyin, S.; Wang, L.; Wusainahong, K.; Ma, C.; Niyazi, M. Comparative study of HPV16 integration in cervical lesions between ethnicities with high and low rates of infection with high-risk HPV and the correlation between integration rate and cervical neoplasia. *Exp. Ther. Med.* **2015**, *10*, 2169–2174. [CrossRef] [PubMed]

48. Liu, L.; Ying, C.; Zhao, Z.; Sui, L.; Zhang, X.; Qian, C.; Wang, Q.; Chen, L.; Guo, Q.; Wu, J. Identification of reliable biomarkers of human papillomavirus 16 methylation in cervical lesions based on integration status using high-resolution melting analysis. *Clin. Epigenetics* **2018**, *10*, 10. [CrossRef] [PubMed]
49. Jiang, Y.; Zhu, C.; He, D.; Gao, Q.; Tian, X.; Ma, X.; Wu, J.; Das, B.C.; Severinov, K.; Hitzeroth, I.I.; et al. Cytological Immunostaining of HMGA2, LRP1B, and TP63 as Potential Biomarkers for Triaging Human Papillomavirus-Positive Women. *Transl. Oncol.* **2019**, *12*, 959–967. [CrossRef]
50. Tuna, M.; Amos, C.I. Next generation sequencing and its applications in HPV-Associated cancers. *Oncotarget* **2017**, *8*, 8877–8889. [CrossRef]
51. Chae, J.; Park, W.S.; Kim, M.J.; Jang, S.S.; Hong, D.; Ryu, J.; Ryu, C.H.; Kim, J.H.; Choi, M.K.; Cho, K.H.; et al. Genomic characterization of clonal evolution during oropharyngeal carcinogenesis driven by human papillomavirus 16. *BMB Rep.* **2018**, *51*, 584–589. [CrossRef]
52. Yang, W.; Liu, Y.; Dong, R.; Liu, J.; Lang, J.; Yang, J.; Wang, W.; Li, J.; Meng, B.; Tian, G. Accurate Detection of HPV Integration Sites in Cervical Cancer Samples Using the Nanopore MinION Sequencer Without Error Correction. *Front. Genet.* **2020**, *11*, 660. [CrossRef] [PubMed]
53. Arroyo-Muhr, L.S.; Lagheden, C.; Hultin, E.; Eklund, C.; Adami, H.O.; Dillner, J.; Sundstrom, K. Human papillomavirus type 16 genomic variation in women with subsequent in situ or invasive cervical cancer: Prospective population-based study. *Br. J. Cancer* **2018**, *119*, 1163–1168. [CrossRef]
54. Blanco, L.; Bernad, A.; Lazaro, J.M.; Martin, G.; Garmendia, C.; Salas, M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **1989**, *264*, 8935–8940. [CrossRef]
55. Binga, E.K.; Lasken, R.S.; Neufeld, J.D. Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology. *ISME J.* **2008**, *2*, 233–241. [CrossRef]
56. Polson, S.W.; Wilhelm, S.W.; Wommack, K.E. Unraveling the viral tapestry (from inside the capsid out). *ISME J.* **2011**, *5*, 165–168. [CrossRef] [PubMed]
57. Li, T.; Unger, E.R.; Rajeevan, M.S. Universal human papillomavirus typing by whole genome sequencing following target enrichment: Evaluation of assay reproducibility and limit of detection. *BMC Genom.* **2019**, *20*, 231. [CrossRef]
58. Tjalma, W. HPV negative cervical cancers and primary HPV screening. *Facts Views Vis. Obgyn* **2018**, *10*, 107–113.
59. Walboomers, J.M.; Jacobs, M.V.; Manos, M.M.; Bosch, F.X.; Kummer, J.A.; Shah, K.V.; Snijders, P.J.; Peto, J.; Meijer, C.J.; Munoz, N. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **1999**, *189*, 12–19. [CrossRef]
60. Lei, J.; Ploner, A.; Lagheden, C.; Eklund, C.; Nordqvist Kleppe, S.; Andrae, B.; Elfstrom, K.M.; Dillner, J.; Sparen, P.; Sundstrom, K. High-risk human papillomavirus status and prognosis in invasive cervical cancer: A nationwide cohort study. *PLoS Med.* **2018**, *15*, e1002666. [CrossRef] [PubMed]
61. De Sanjose, S.; Quint, W.G.; Alemany, L.; Geraets, D.T.; Klaustermeier, J.E.; Lloveras, B.; Tous, S.; Felix, A.; Bravo, L.E.; Shin, H.R.; et al. Human papillomavirus genotype attribution in invasive cervical cancer: A retrospective cross-sectional worldwide study. *Lancet Oncol.* **2010**, *11*, 1048–1056. [CrossRef]
62. Arroyo Muhr, L.S.; Lagheden, C.; Eklund, C.; Lei, J.; Nordqvist-Kleppe, S.; Sparen, P.; Sundstrom, K.; Dillner, J. Sequencing detects human papillomavirus in some apparently HPV-negative invasive cervical cancers. *J. Gen. Virol.* **2020**, *101*, 265–270. [CrossRef]
63. Gillison, M.L.; D'Souza, G.; Westra, W.; Sugar, E.; Xiao, W.; Begum, S.; Viscidi, R. Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. *J. Natl. Cancer Inst.* **2008**, *100*, 407–420. [CrossRef]
64. Wakeham, K.; Kavanagh, K.; Cuschieri, K.; Millan, D.; Pollock, K.G.; Bell, S.; Burton, K.; Reed, N.S.; Graham, S.V. HPV status and favourable outcome in vulvar squamous cancer. *Int. J. Cancer* **2017**, *140*, 1134–1146. [CrossRef] [PubMed]
65. O'Sullivan, B.; Huang, S.H.; Su, J.; Garden, A.S.; Sturgis, E.M.; Dahlstrom, K.; Lee, N.; Riaz, N.; Pei, X.; Koyfman, S.A.; et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): A multicentre cohort study. *Lancet Oncol.* **2016**, *17*, 440–451. [CrossRef]
66. World Health Organization. Female Genital Tumors. *WHO Classification of Tumors*, 5th ed. Volume 5. Available online: https://publications.iarc.fr/592 (accessed on 4 April 2021).
67. Lubock, N.B.; Zhang, D.; Sidore, A.M.; Church, G.M.; Kosuri, S. A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Res.* **2017**, *45*, 9206–9217. [CrossRef]
68. Mitchell, K.; Brito, J.J.; Mandric, I.; Wu, Q.; Knyazev, S.; Chang, S.; Martin, L.S.; Karlsberg, A.; Gerasimov, E.; Littman, R.; et al. Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biol.* **2020**, *21*, 71. [CrossRef] [PubMed]
69. Lindgreen, S. AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **2012**, *5*, 337. [CrossRef]
70. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194. [CrossRef] [PubMed]
71. Broad Institute. Genome Analysis Toolkit. Available online: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores (accessed on 26 May 2021).
72. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.20133997.
73. Li, R.; Yu, C.; Li, Y.; Lam, T.W.; Yiu, S.M.; Kristiansen, K.; Wang, J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966–1967. [CrossRef]

74. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
75. Novoalign. Available online: http://novocraft.com/ (accessed on 26 May 2021).
76. Sedlazeck, F.J.; Rescheneder, P.; Von Haeseler, A. NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **2013**, *29*, 2790–2791. [CrossRef]
77. Arroyo Muhr, L.S.; Eklund, C.; Dillner, J. Misclassifications in human papillomavirus databases. *Virology* **2021**, *558*, 57–66. [CrossRef] [PubMed]
78. Ekstrom, J.; Muhr, L.S.; Bzhalava, D.; Soderlund-Strand, A.; Hultin, E.; Nordin, P.; Stenquist, B.; Paoli, J.; Forslund, O.; Dillner, J. Diversity of human papillomaviruses in skin lesions. *Virology* **2013**, *447*, 300–311. [CrossRef]
79. de Villiers, E.M.; Fauquet, C.; Broker, T.R.; Bernard, H.U.; zur Hausen, H. Classification of papillomaviruses. *Virology* **2004**, *324*, 17–27. [CrossRef] [PubMed]
80. Bernard, H.U.; Burk, R.D.; Chen, Z.; Van Doorslaer, K.; Zur Hausen, H.; De Villiers, E.M. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **2010**, *401*, 70–79. [CrossRef]
81. De Villiers, E.M. Cross-roads in the classification of papillomaviruses. *Virology* **2013**, *445*, 2–10. [CrossRef] [PubMed]
82. Burk, R.D.; Harari, A.; Chen, Z. Human papillomavirus genome variants. *Virology* **2013**, *445*, 232–243. [CrossRef] [PubMed]
83. Smith, B.; Chen, Z.; Reimers, L.; Van Doorslaer, K.; Schiffman, M.; Desalle, R.; Herrero, R.; Yu, K.; Wacholder, S.; Wang, T.; et al. Sequence imputation of HPV16 genomes for genetic association studies. *PLoS ONE* **2011**, *6*, e21375. [CrossRef]
84. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
85. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
86. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
87. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
88. International Journal of Cancer. Submission Guidelines. Available online: https://onlinelibrary.wiley.com/pb-assets/assets/10970215/IJC_Sequencing_Coverage_and_Quality_Statistics_Guidelines-1607431877843.pdf (accessed on 26 May 2021).
89. Poljak, M.; Cuschieri, K.; Waheed, D.E.; Baay, M.; Vorsters, A. Impact of the COVID-19 pandemic on human papillomavirus-based testing services to support cervical cancer screening. *Acta Dermatovenerol. Alp. Pannonica Adriat.* **2021**, *30*, 21–26.
90. Hilke, F.J.; Muyas, F.; Admard, J.; Kootz, B.; Nann, D.; Welz, S.; Riess, O.; Zips, D.; Ossowski, S.; Schroeder, C.; et al. Dynamics of cell-free tumour DNA correlate with treatment response of head and neck cancer patients receiving radiochemotherapy. *Radiother. Oncol.* **2020**, *151*, 182–189. [CrossRef]
91. Wagner, S.; Roberson, D.; Boland, J.; Yeager, M.; Cullen, M.; Mirabello, L.; Dunn, S.T.; Walker, J.; Zuna, R.; Schiffman, M.; et al. Development of the TypeSeq Assay for Detection of 51 Human Papillomavirus Genotypes by Next-Generation Sequencing. *J. Clin. Microbiol.* **2019**, *57*. [CrossRef] [PubMed]
92. Chan, W.S.; Chan, T.L.; Au, C.H.; Leung, C.P.; To, M.Y.; Ng, M.K.; Leung, S.M.; Chan, M.K.M.; Ma, E.S.K.; Tang, B.S.F. An economical Nanopore sequencing assay for human papillomavirus (HPV) genotyping. *Diagn. Pathol.* **2020**, *15*, 45. [CrossRef]
93. Lahens, N.F.; Ricciotti, E.; Smirnova, O.; Toorens, E.; Kim, E.J.; Baruzzo, G.; Hayer, K.E.; Ganguly, T.; Schug, J.; Grant, G.R. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genom.* **2017**, *18*, 602. [CrossRef]
94. Marine, R.L.; Magana, L.C.; Castro, C.J.; Zhao, K.; Montmayeur, A.M.; Schmidt, A.; Diez-Valcarce, M.; Ng, T.F.F.; Vinje, J.; Burns, C.C.; et al. Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. *J. Virol. Methods* **2020**, *280*, 113865. [CrossRef]
95. Hornung, B.V.H.; Zwittink, R.D.; Kuijper, E.J. Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* **2019**, *95*. [CrossRef] [PubMed]
96. Zur Hausen, H. Papillomaviruses and cancer: From basic studies to clinical application. *Nat. Rev. Cancer* **2002**, *2*, 342–350. [CrossRef] [PubMed]
97. Bzhalava, Z.; Arroyo Muhr, L.S.; Dillner, J. Transcription of human papillomavirus oncogenes in head and neck squamous cell carcinomas. *Vaccine* **2020**, *38*, 4066–4070. [CrossRef] [PubMed]
98. Braakhuis, B.J.; Snijders, P.J.; Keune, W.J.; Meijer, C.J.; Ruijter-Schippers, H.J.; Leemans, C.R.; Brakenhoff, R.H. Genetic patterns in head and neck cancers that contain or lack transcriptionally active human papillomavirus. *J. Natl. Cancer Inst.* **2004**, *96*, 998–1006. [CrossRef]
99. Leemans, C.R.; Braakhuis, B.J.; Brakenhoff, R.H. The molecular biology of head and neck cancer. *Nat. Rev. Cancer* **2011**, *11*, 9–22. [CrossRef]
100. World Health Organization. Available online: http://www.who.int/biologicals/vaccines/hpv/en/index.htm (accessed on 26 May 2021).
101. World Health Organization. *Human Papillomavirus Laboratory Manual*, 1st ed.; Geneva, World Health Organization: Geneva, Switzerland, 2009. Available online: https://apps.who.int/iris/handle/10665/70505 (accessed on 4 April 2021).

102. Eklund, C.; Forslund, O.; Wallin, K.L.; Dillner, J. Continuing global improvement in human papillomavirus DNA genotyping services: The 2013 and 2014 HPV LabNet international proficiency studies. *J. Clin. Virol.* **2018**, *101*, 74–85. [CrossRef] [PubMed]
103. Cuschieri, K.; Schuurman, R.; Coughlan, S. Ensuring quality in cervical screening programmes based on molecular human papillomavirus testing. *Cytopathology* **2019**, *30*, 273–280. [CrossRef] [PubMed]
104. Fagan, E.J.; Moore, C.; Jenkins, C.; Rossouw, A.; Cubie, H.A.; James, V.L. External quality assessment for molecular detection of human papillomaviruses. *J. Clin. Virol.* **2010**, *48*, 251–254. [CrossRef] [PubMed]