THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale ED 340
BIOLOGIE MOLÉCULAIRE INTÉGRATIVE ET CELLULAIRE (BMIC)

Spécialité de doctorat : Biologie moléculaire

Soutenue publiquement le 15/10/2020, par :

Rosario Nicola Brancaccio

Novel strategies for the identification and full genomic characterization of unknown HPV types from human DNA samples

---

Devant le jury composé de :

Bernet Agnès, Professeure des Universités, Université de Lyon 1	Examinatrice
Bravo Ignacio, Directeur de Recherche, CNRS de Montpellier	Rapporteur
Steenbergen Renske, Professeure Associée, Université de Amsterdam	Rapporteure
Chemin Isabelle, Directrice de Recherche, Université de Lyon	Examinatrice
Ottonello Simone, Professeur des Universités, Université de Parme	Examinateur
Tommasino Massimo, Directeur de Recherche/Chef de Section, IARC	Directeur de thèse
Gouy Manolo, Directeur de Recherche, CNRS de Lyon	Invité

Laboratory

Infection and Cancer Biology Group (ICB)
Section of Infections (INF)
International Agency for Research on Cancer (IARC)
World Health Organization (WHO)
150, cours Albert Thomas
69372 Lyon CEDEX 08
FRANCE

Résumé

Les papillomavirus humains (HPV) sont de petits virus icosaédriques non-enveloppés à ADN circulaire double brin. À ce jour, plus de 200 HPV ont été identifiés. La classification des Papillomaviridae est basée sur les identités de séquence nucléotidique du gène L1.

La découverte de nouveaux HPV est d'une importance cruciale pour élargir nos connaissances sur ces virus. Alors qu'un sous-groupe HPV du genre alpha communément appelé "à haut risque", a été clairement associé aux cancers anogénitaux, et à certains cancers de la tête et du cou, il existe peu d'informations sur le rôle des HPV des genres bêta et gamma-HPVs dans les cancers humains. Des études récentes montrent que des bêta-HPVs pourraient jouer un rôle dans les cancers de la peau de type non-mélanome (NMSC), en association avec le rayonnement ultraviolet (UV). Toutefois, des études supplémentaires sont nécessaires pour confirmer cette hypothèse chez l'homme.

Jusqu'à présent, les méthodes de choix utilisées pour l'identification de nouveaux HPVs dans des échantillons humains reposaient sur l'utilisation de la PCR et le séquençage de Sanger. Plusieurs amorces spécifiques ou à large spectre ciblant le gène L1, ont été développées pour l'amplification de séquences de HPV. De nos jours, avec l'avènement des méthodes de séquençage à haut débit (NGS), la possibilité d'élargir ces familles virales s'est accrue. Cependant, de grandes quantités de données sont générées par ces nouvelles méthodes de séquençage nécessitant le développement de méthodes bioinformatiques d'analyse spécifiques pour une identification et une reconstruction efficaces des séquences virales.

Le but de ce travail de thèse a été de développer une stratégie pour l'identification des HPV, ainsi que la caractérisation du génome complet de nouveaux types, dans des échantillons humains. Au cours de ce travail de thèse, l'utilisation d'amorces à large spectre connues et validées, mais aussi l'utilisation d'amorces améliorées par notre laboratoire, toutes ciblant une portion du gène L1, ont été utilisées pour la recherche de séquences HPV dans des échantillons cutanés et oraux. L'analyse des données NGS a permis l'identification de 105 possible nouvelles séquences de PVs, ainsi que l'identification de 296 HPV connus. Une méthode d'analyse bioinformatique a été développée pour l'analyse des données de séquençage.

La deuxième partie de ma thèse a consisté, à partir des données NGS précédemment obtenues, d'obtenir le génome entier de plusieurs HPVs. Ainsi, le génome entier de HPV227, un nouveau bêta-HPV, a été obtenu, en combinant plusieurs méthodes c.à.d. l'amplification par cercle roulant (ou rolling circle amplification - RCA), PCR pour amplification de longs fragments (ou "Long Range PCR"), le séquençage de Sanger par "primer walking", et le clonage.

Enfin, la technologie de séquençage par nanopores (MinION, Oxford Nanopore Technologies) a été utilisée afin d'obtenir le génome entier du HPV227, directement à partir de l'échantillon cutané d'origine et ainsi permettre un gain de temps important dans la caractérisation du génome complet des HPV. Pour ce faire, une méthode d'analyse bioinformatique des données de séquençage MinION pour la reconstruction des génomes de HPV a été développée. L'ensemble du génome du HPV227 a été reconstruit, confirmant l'efficacité de cette stratégie.

Dans l'ensemble, cette thèse décrit et valide plusieurs approches pour l'identification et la caractérisation complète de nouveaux génomes de HPV. De plus, la découverte de 105 possible nouveaux PVs élargit nos connaissances sur cette famille de virus, bien que pour des analyses supplémentaires soient nécessaires pour la caractérisation complète de ces nouveaux virus.

Abstract

Human papillomaviruses (HPV) are small non-enveloped icosahedral viruses with double-stranded circular DNA. More than 200 different human papillomaviruses have been identified so far. The classification of the *papillomaviridae* family is based on pairwise nucleotide sequence identity across the L1 ORF, as it is well conserved and can allowa genome-based approach to PV classification.

The discovery of new HPVs is of paramount importance to expand our knowledge on the role of these viruses in human diseases. In particular, while a subgroup of alphapapillomaviruses, referred to as high-risk HPV types, have been related to anogenital cancer and a subset of head and neck cancers, less is known about the role of HPVs from the other genera, such as beta and gamma types, in human cancers. Recent studies show a potential role of betapapillomaviruses in cutaneous squamous cell carcinoma (cSCC) together with ultraviolet (UV) radiation, but further studies are required. The methods used so far for the identification of novel HPVs in human specimens are based on PCR and Sanger sequencing, and several specific and broad-spectrum PCR primers, targeting the L1 ORF, have been designed for the amplification of HPV sequences. Nowadays, with the advent of high-throughput sequencing methods (i.e., NGS, ONT), we can expand this viral family with a more straightforward approach. Additionally, the large amount of data, generated by these new sequencing methods, requires the development of specific bioinformatics analyses for the identification and reconstruction of the viral sequences.

The aim of this work was the development of an effective strategy for the identification and full genomic characterization of novel HPV types in human samples.

In this work, known and new broad-spectrum PCR primers, all targeting a portion of the HPV L1 ORF, were used to amplify the HPV sequence on human skin (n=119) and oral (n=147) samples from healthy individuals. After, NGS analysis allowed the identification of 105 putative novel HPV types in addition to 296 known types. A specific workflow was developed to analyze the NGS sequencing data. In the second step of this work, starting from the NGS data, the whole genome of HPV227, a novel beta-2 papillomavirus, was obtained using a primer-walking strategy and Sanger sequencing. Finally, the MinION sequencing technology was used to obtain the whole genome of HPV227, directly from the original skin sample where this virus was discovered. A bioinformatics analysis was developed for the reconstruction of HPV genomes using MinION sequencing data. The entire genome of HPV227 was reconstructed, confirming the effectiveness of this strategy. Overall, this thesis describes a valid approach for the identification and full characterization of novel HPV genomes. Moreover, the discovery of 105 putative novel PV types expands our knowledge on this family of viruses, although further analyses are required for a complete characterization of these new viruses.

Résumé en français

Les papillomavirus humains (HPV) sont de petits virus icosaédriques non-enveloppés à ADN circulaire double brin qui infectent les épithéliums cutanés et les muqueuses. À ce jour, plus de 200 HPV ont été identifiés, et classés en 5 genres phylogénétiques : alpha, bêta, gamma, mu et nu. Un sous-groupe HPV du genre alpha communément appelé "à haut risque" ou "HPV HR", a été clairement associé aux cancers anogénitaux, et à certains cancers de la tête et du cou. Les HPV cutanés, principalement issus du genre bêta ou gamma, se retrouvent fréquemment à la surface de la peau de la population générale. Cependant, plusieurs études montrent le rôle potentiel des bêta-papillomavirus dans le cancer de la peau de type non mélanome, en association avec le rayonnement ultraviolet (UV). Les HPV HR induisent les cancers, notamment du col de l'utérus, en modifiant les voies impliquées dans la réponse immunitaire de l'hôte, afin d'établir une infection persistante et favoriser la transformation cellulaire, par l'expression constitutive des oncoprotéines virales E6 et E7. Dans le cas des HPV cutanés, l'expression de E6 et E7 n'est pas nécessaire au maintien du phénotype du cancer de la peau, et un mécanisme "hit and run" est proposé pour expliquer le rôle potentiel des HPV cutanés dans l'initiation de la cancérogenèse cutanée en agissant comme des facilitateurs plutôt que comme acteurs directs dans la carcinogenèse. Toutefois, des études supplémentaires sont nécessaires pour confirmer cette hypothèse chez l'homme.

La classification des Papillomaviridae est basée sur l'homologie des séquences nucléotidiques du cadre de lecture L1. Les différents genres partagent moins du 60% d'identité nucléotidique. Les différentes espèces d'un même genre partagent entre 60% et 70% d'identité nucléotidique. Les HPV appartenant à la même espèce partagent entre 71% et 89% d'identité nucléotidique. De plus, les PV partageant entre 90% et 98% d'identité nucléotidique appartiennent au même sous-type, tandis que ceux qui partagent plus de 98% d'identité nucléotidique sont considérés comme des variants.

La découverte de nouveaux HPV est d'une importance capitale pour mieux appréhender le rôle de ces virus dans les maladies humaines. Les méthodes utilisées jusqu'à présent pour l'identification de nouveaux HPV, sont basées sur la technique de la PCR avec l'utilisation d'amorces spécifiques et à large spectre ciblant le cadre de lecture L1, et le séquençage de Sanger. Avec l'avènement des méthodes de séquençage à haut débit (i.e. NGS, ONT), cette famille virale s'étoffe de façon exponentielle avec la découverte récente d'un nombre important de HPV. Cependant, la grande quantité de données générées par ces nouvelles méthodes de séquençage nécessite le développement de méthodes bioinformatiques d'analyse spécifiques pour une identification et une reconstruction efficaces des séquences virales.

Le but de ce travail de thèse a été de développer une stratégie pour l'identification des HPV, ainsi que la caractérisation du génome complet de nouveaux types, dans des échantillons humains. Au cours de ce travail de thèse, l'utilisation d'amorces à large spectre connues et validées, mais aussi l'utilisation d'amorces améliorées par notre laboratoire, toutes ciblant une portion du gène L1, ont été utilisées pour la recherche de séquences HPV dans des échantillons cutanés et oraux. Les produits PCR ont ensuite été purifiés et regroupés avant d'être séquencé. L'analyse des données NGS a permis l'identification de 105 possible nouvelles séquences de PV, ainsi que l'identification de 296 HPV connus. Une méthode d'analyse bioinformatique, PVAmpliconFinder, a été développée pour l'analyse de ces données de séquençage, et optimisée pour identifier les types HPV potentiellement nouveaux.

Le second objectif de ma thèse a consisté à caractériser le génome entier des nouveaux HPV. Ainsi, à partir de séquences partielles du gène L1 (i.e. 99 pb représentant un nouveau type HPV, découvert à partir des analyses des données NGS, et nommé HPV-ICB2), le génome complet de nouveaux types HPV ont pu être obtenu grâce à l'utilisation de l'amplification par cercle roulant (ou rolling circle amplification - RCA), l'utilisation de la PCR pour l'amplification de longs

fragments (ou “Long Range PCR”), le séquençage de Sanger par “primer walking”, et le clonage. Le nom officiel «HPV227» a été attribué à HPV-ICB2.
Enfin, le troisième objectif de ma thèse a consisté à utiliser la technologie de séquençage par nanopores (MinION, Oxford Nanopore Technologies). Cette technologie a été évaluée et validée en séquençant le génome entier de HPV227 précédemment caractérisé par le séquençage de Sanger. L’utilisation de cette technologie, directement à partir de l’échantillon d’origine, a permis un gain de temps important dans la caractérisation du génome complet de HPV. Une méthode d’analyse bioinformatique des données de séquençage MinION pour la reconstruction des génomes de HPV a été spécifiquement développée. L'ensemble du génome du HPV227 a été reconstruit avec plus de 99.9% d’identité, confirmant l'efficacité de cette stratégie.
Dans l'ensemble, cette thèse décrit plusieurs stratégies efficaces pour l'identification et la caractérisation complète de nouveaux génomes de HPV. De plus, la découverte de 105 nouveaux types de PV potentiellement humains permet d’élargir nos connaissances sur cette famille de virus.

Acknowledgements

I wish to express my sincere appreciation to my supervisors, Dr Massimo Tommasino and Dr Tarik Gheit. During these years of PhD, they convincingly guided and encouraged me to be professional and do the right thing even when the road got tough. Without their persistent help, the goal of this project would not have been realized. I will treasure their teachings.

I wish to thank all the colleagues of the ICB group for their support and collaboration during these years. Their help made the difference.

I wish to acknowledge Nicole Suty for her constant help in many different aspects of my job.

I want to thank also to all the colleagues from the other IARC's groups. Their help was essential in many critical points of my job. In particular, I want to acknowledge Cyrille Cuenin that was always friendly and qualified.

I am also grateful to all the ITS members for their help. In particular, I want to thank Nicolas Tardy for his competence and kindness.

I wish to thank the members of my PhD jury for their contribution and their extreme availability.

Finally, I wish to acknowledge the support and great love of my family and my fiancée, Emanuela. They kept me going on and this work would not have been possible without their input.

List of acronyms

AE early polyadenylated sites

AGC Atypical Glandular Cells

AGW Anogenital warts

AIDS Acquired Immunodeficiency Syndrome

AIN Anal Intraepithelial Neoplasia

AK Actinic Keratoses

AL late polyadenylated sites

ART Antiretroviral Therapies

ASC Atypical squamous cells

ASIR infection-attributable age-standardized incidence rate

ATP Adenosine Tri-Phosphate

BCC Basal Cell Carcinomas

BLAST Basic Local Alignment Search Tool

CA Condyloma Acuminatum

cds Coding DNA Sequence

CE Capillary Electrophoresis

CIN Cervical Intraepithelial Neoplasia

CKII casein kinase II

CMOS Metal-Oxide-Semiconductor

CODEHOP Consensus-degenerate hybrid oligonucleotide primers

CpG 5'-C-phosphate-G-3'

cSCC Cutaneous Squamous Cell Carcinoma

DNA Deoxyribonucleic Acid

DOCK8 Dedicator of Cytokinesis 8

dsDNA Double Strand Deoxyribonucleic Acid

EBV Epstein-Barr virus

ECM extracellular matrix

EPA Evolutionary Placement Algorithm

ETQL Expression quantitative trait loci

EV Epidermodysplasia verruciformis

FADD Fas-associated death domain

FM indexes Ferragina-Manzini indexes

GTR General Time-Reversible model

GWAS Genome-Wide Association Studies

HAC High Accuracy Calling

HBV hepatitis virus B

HCV hepatitis virus C

HIV Human Immunodeficiency Virus

HLTV-1 human T lymphotropic virus-1

HNSCC Head and neck squamous cell carcinoma

HPV Human Papillomaviruses

HR High Risk

HRA High-Resolution Anoscopy

HSCT Hematopoietic Stem Cell Transplantation

HSCT Hematopoietic Stem Cell Transplantation

HSIL High-grade Squamous Intraepithelial Lesion

HTML Hypertext Markup Language

IAP inhibitor of apoptosis

IARC International Agency for Research on Cancer

ICB Infection and Cancer Biology Group

ICTV International Committee for the Taxonomy of Viruses

ID Identifier

IHC Immunohistochemistry

INF Section of Infections

kb Kilobase

kDa Kilo Dalton

KHSV Kaposi's Sarcoma-associated virus

KSHV Kaposi's sarcoma-associated herpesvirus

LCA Last Common Ancestor

LCR Long Control Region

LR Low Risk

LSIL Low-grade Squamous Intraepithelial Lesion

MCV Merkel cell polyomavirus

MDA Multiple Displacement Amplification

ML Maximum Likelihood

MSM men who have sex with men

MUSCLE MUltiple Sequence Comparison by Log-Expectation

NCBI National Center for Biotechnology Information

NCOR nuclear receptor corepressor

NGS Next Generation Sequencing

NMSC Non-Melanoma Skin Cancer

ONT Oxford Nanopore Technologies

ORF Open Reading Frame

pA polyadenylated sites

PaVE Papillomavirus Episteme

PCR Polymerase Chain Reaction

PNPV National Vaccine Prevention Plan

PPV Positive Predictive Value

PV Papillomaviruses

QC Quality Control

qPCR Quantitative Polymerase Chain Reaction

RaxML Randomized Axelerated Maximum Likelihood

RCA Rolling circle amplification

RNA Ribonucleic Acid

rRNA Ribosomal Ribonucleic Acid

RRP Recurrent Respiratory Papillomatosis

S phase Synthesis Phase

SBS Sequence-By-Synthesis

SCC Squamous Cell Carcinoma

SMRT silencing mediator of retinoic acid and thyroid hormone receptor

SMRT Single-Molecule Real-Time

SMS Single Molecule Sequencing

SNP Single Nucleotide Polymorphism

SOT Solid Organ Transplant

SRA Sequence Read Archive

URR Upstream Regulatory Region

USA United States of America

UTR Untranslated Region

UV Ultraviolet

UVR Ultraviolet Radiation

WGA Whole Genome Amplification

WGS Whole Genome Sequencing

WHO World Health Organization

Table of Contents

Introduction

1 Role of infectious agents in cancer development

The first evidence of the existence of viruses came from the experiment of Ivanovski, who allowed the discovery in 1892 of the tobacco mosaic virus[1], while the primary hypothesis that animal virus could play a role in cancer formation was formulated by Loeffler and Frosch in 1898[2]. In 1901, Reed identified the yellow fever virus in human[3]. A few years later, viruses were proposed as a common cause of human cancers.

The first tumor virus, Rous Sarcoma virus that causes sarcoma in chickens, was discovered by Peyton Rous from Rockefeller Institute in 1911[4,5].

In 1964, Epstein, Achong and Barr, identified Epstein-Barr virus (EBV), the first human tumor virus, from Burkitt lymphoma cell line by using electron microscopy[6].

In 1984, Warren and Marshall discovered Helicobacter pylori, a bacterium associated with gastritis and peptic ulceration[7,8].

In 1970 Harald Zur Hausen hypothesized the relationship between HPV and cervical cancer, for which he received the Nobel Prize in Physiology or Medicine 2008[9].

It is also worth mentioning the Kaposi's Sarcoma-associated virus (KHSV) discovered by Yuan Chang in 1994, and the Merkel cell polyomavirus (MCV) found in Merkel cell carcinoma, by Patrick Moore in 2008[10–12] (Figure 1).

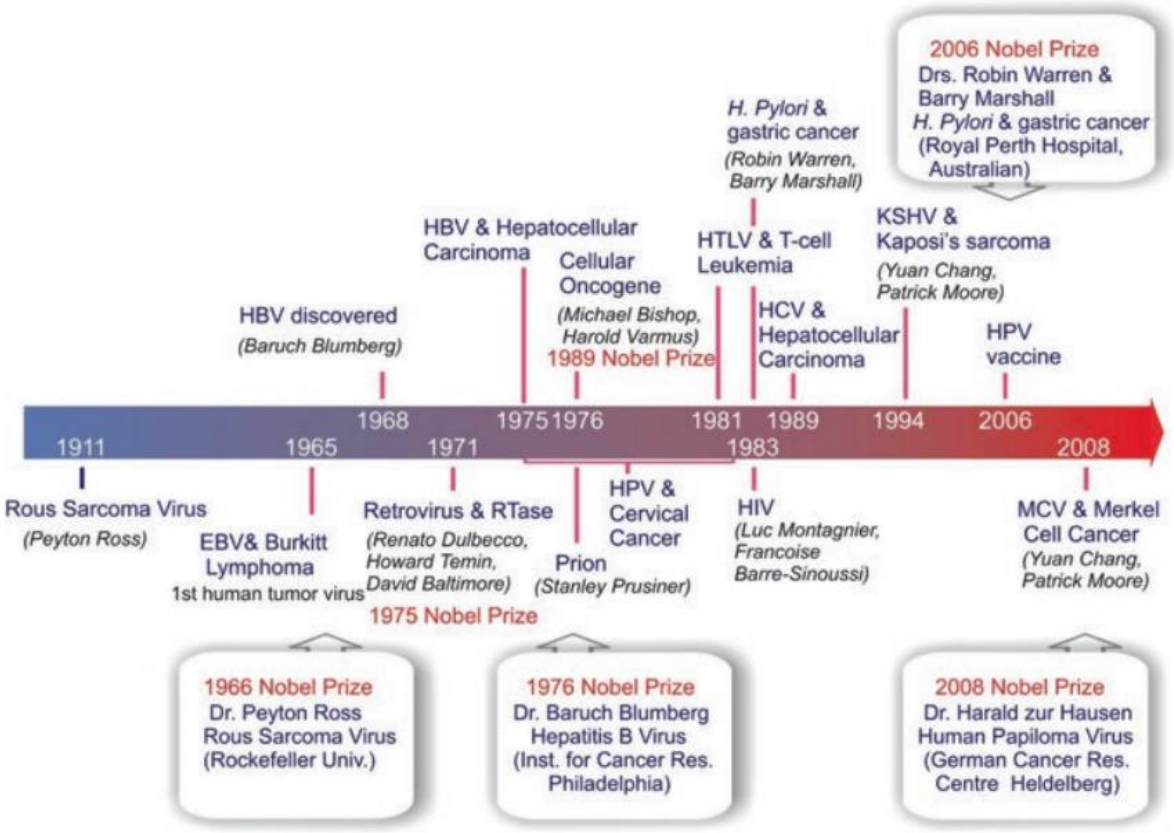


Figure 1: Most relevant discoveries in history that led to understanding the role of infectious agents in human cancer development [2]

It has been established that, during their lifetime, 99% of the population is infected with organisms known to cause cancers[13].

Different infectious agents can cause cancers, particularly in less developed countries[13,14].

The global burden of infections-related cancers is a topic of particular interest studied to understand the mechanisms of cancer development and prevention. Several studies assessed the contribution of infectious agents in cancer development[13,15–17].

A total of 14 million new cancer cases worldwide are diagnosed every year (Table 1)[18].

In 2018, an estimated 2.2 million infection-attributable cancer cases were diagnosed worldwide, corresponding to an infection-attributable age-standardized incidence rate (ASIR) of 25 cases per 100,000 person-years[17].

Sixty-six percent of all the infection-attributable cancers occur in developing countries[18]. This percentage is lower than 5% in the USA, Canada, Australia, New Zealand, and some countries in northern and western Europe, but higher than 40% in many countries in sub-Saharan Africa and Mongolia[18] (Table 1).

In general, in more developed regions, the percentage of cancer cases attributed to infections is lower than 10% while in less developed countries, it exceeds the 23%[18].

	Number of new cases	Number attributable to infection	Attributable fraction (%)
Worldwide	14 000 000	2 200 000	15·4%
Africa			
Sub-Saharan Africa	630 000	200 000	31·3%
North Africa and west Asia	540 000	70 000	13·1%
Asia			
Central Asia	1 500 000	290 000	19·4%
East Asia	4 900 000	1 100 000	22·8%
America			
Latin America	1 100 000	160 000	14·4%
North America	1 800 000	72 000	4·0%
Europe	3 400 000	250 000	7·2%
Oceania	160 000	7600	4·9%
Human Development Index			
Very high	5 700 000	430 000	7·6%
High	2 200 000	290 000	13·2%
Medium	5 200 000	1 200 000	23·0%
Low	940 000	240 000	25·3%
Level of development			
More developed regions	7 900 000	730 000	9·2%
Less developed regions	6 200 000	1 400 000	23·4%

Numbers of cases rounded to two significant figures.

Table 1: Number of new cancer cases associated with infectious agents, by geographical region and level of development[18]

More than 90% of infection-related cancers are due to viral infections[2] including human papillomaviruses (HPV 16 and 18)[19], hepatitis viruses (HBV and HCV)[20], Epstein-Barr virus (EBV)[21], Kaposi's sarcoma-associated herpesvirus (KSHV)[22], human T lymphotropic virus-1(HLTV-1)[23] and Merkel cell polyomavirus (MCV)[24].

In addition to viruses causing cancers, there are also other non-viral infections like the bacterium *Helicobacter pylori* [25] and some parasites like *Schistosoma haematobium*[26] that are known to be involved in cancer development[2].

In an epidemiological study published in 2020, Catherine de Martel and colleagues evaluated the incidence of cancers attributable to infectious agents. *Helicobacter pylori* was identified as the primary cause of infections-related cancers (810,000 cases, ASIR 8.7 cases per 100 000 person-years), followed by HPV (690,000 cases, ASIR 8.0), HBV (360,000 cases, ASIR 4.1) and HCV (160,000 cases, ASIR 1.7). Additional 210,000 new cases were attributed to Epstein-Barr virus, human herpesvirus type 8 (HHV8; also known as Kaposi sarcoma herpesvirus), human T-cell lymphotropic virus type 1 (HTLV-1), and parasitic infections (i.e., *Schistosoma haematobium, Opisthorchis viverrini* and *Clonorchis sinensis*) (Table 2)[17].

In men, the leading cause of infection-related cancers in 2018 was *Helicobacter pylori,* with 525.700 new cases*,* followed by HBV (270,000 new cases), HCV (108,700 new cases), Epstein-Barr virus (104,100 new cases) and HPV (69,400 new cases). In women, HPV represents the first cause of infection-related cancer with more than 620,000 new cases in 2018, followed by *Helicobacter pylori* (286,500), HBV (90,000), Epstein-Barr virus (48,500) and HCV (47,200), (Table 2)[17].

There is also a relation between cancer incidence and age. Overall, a significant percentage of cancers, i.e., 64%, occur before age 70 years[27], and regarding those caused by infectious agents, this percentage is even higher both in less developed and more developed countries[18].

For *Helicobacter pylori,* a higher infection-related cancer incidence is registered after age 50, while for HPV the 86% of cases occur before age 70[18].

	Men		Women		Total	
	New cases	New cases attributable to infectious pathogens	New cases	New cases attributable to infectious pathogens	New cases	New cases attributable to infectious pathogens
Helicobacter pylori						
Non-cardia gastric cancer	550 000	490 000	300 000	270 000	850 000	760 000
Cardia gastric cancer	130 000	27 000	46 000	8900	180 000	36 000
Non-Hodgkin lymphoma of gastric location	12 000	8700	10 000	7600	22 000	16 000
Human papillomavirus						
Cervix uteri carcinoma*	..	..	570 000	570 000	570 000	570 000
Oropharyngeal carcinoma	110 000	34 000	26 000	8100	140 000	42 000
Oral cavity cancer	190 000	3900	91 000	2000	280 000	5900
Larynx cancer*	150 000	3600	22 000	≤1000	180 000	4100
Anus squamous cell carcinoma	9900	9900	19 000	19 000	29 000	29 000
Penis carcinoma*	34 000	18 000	..	..	34 000	18 000
Vagina carcinoma*	..	..	18 000	14 000	18 000	14 000
Vulva carcinoma*	..	..	44 000	11 000	44 000	11 000
Hepatitis B virus						
Hepatocellular carcinoma	490 000	270 000	170 000	90 000	660 000	360 000
Hepatitis C virus						
Hepatocellular carcinoma	490 000	100 000	170 000	40 000	660 000	140 000
Other non-Hodgkin lymphoma	260 000	8700	210 000	7200	480 000	16 000
Epstein-Barr virus						
Nasopharynx carcinoma*	92 000	76 000	35 000	29 000	130 000	110 000
Hodgkin lymphoma*	46 000	24 000	33 000	17 000	80 000	40 000
Burkitt lymphoma	7800	4100	3800	2500	12 000	6600
Human herpesvirus type 8						
Kaposi sarcoma*	28 000	28 000	14 000	14 000	42 000	42 000
Schistosoma haematobium						
Bladder carcinoma*	420 000	4000	120 000	1900	550 000	6000
Human T-cell lymphotropic virus						
Adult T-cell leukaemia and lymphoma	1900	1900	1700	1700	3600	3600
Opisthorchis viverrini and *Clonorchis sinensis*						
Cholangiocarcinoma	69 000	2100	56 000	1300	130 000	3500
All cancer types related to infection	..	1 100 000	..	1 100 000	..	2 200 000

The number of cases has been rounded to two significant digits. *Cancer site for which estimates were available in, and extracted directly from, GLOBOCAN 2018 via the Cancer Today website.

Table 1: **Estimated numbers of infection-attributable cancer cases in 2018, by infectious pathogen, cancer subsite, and sex**

Table 2: Number of new cancer cases in 2018 attributable to infection, by infectious agent, cancer subsite, and sex[17]

Stomach-, liver-, and cervix- cancers are the most frequent cancers worldwide associated with infectious agents[18], representing the fifth, sixth, and seventh-most common cancers worldwide, respectively[28].

These cancer types also have very high infection-related proportions[18]. All cervical cancers are associated with HPV[18]. In 2012, 530,000 cervical-cancer cases worldwide had been attributed to HPV infections, while only 113,000 cancer cases other than cervical cancer cases had been assigned to HPV[18]. In 2012, in both less and more developed countries, the attributable population fractions derived from infection-related cancers were generally higher in younger people, between 40 and 45 years old. For women in more developed countries, these values were higher in people younger than 40 years (Figure 2)[18].

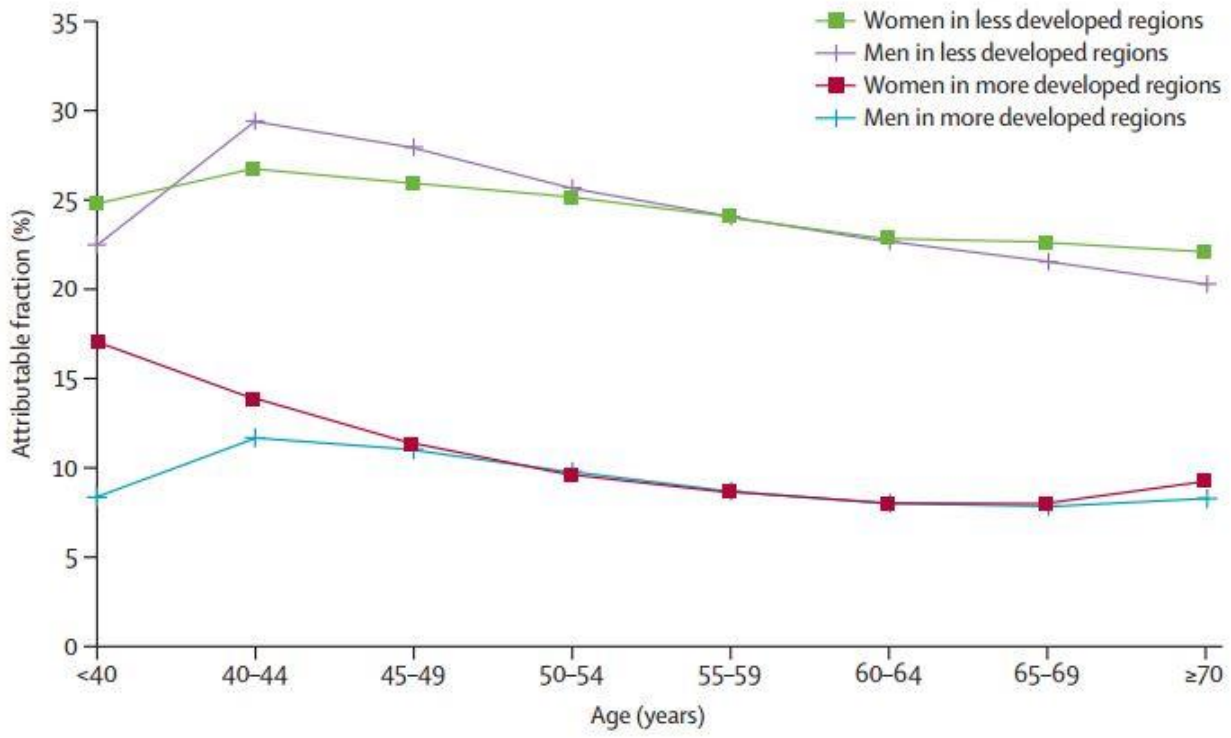


Figure 2: Representation of the relation of the percentages of cancer cases in 2012 attributable to infection with sex, age group, and development status[18]

2 Vaccination programs for the prevention of the HPV-related cancers

Human papillomavirus infection-related cancers include those of the cervix, vulva, vagina, penis, anus, rectum, and oropharynx[29]. Over 80% of anogenital cancers occur in the cervix. Several studies aim to develop effective vaccines for the prevention of HPV-related cancers[30].

Three vaccines are available nowadays (i.e., Gardasil, Gardasil9, and Cervarix), all containing synthetically manufactured virus-like particles (VLPs) of the L1 epitope (Table 3). All three vaccines prevent infection with HPV types 16 and 18[31,32].

Gardasil offers additional protection against genotypes 6 and 11, causing over 90% of anogenital warts and virtually all cases of Recurrent Respiratory Papillomatosis (RRP) in both sexes[33–35]. Gardasil 9 prevents infection with the same four HPV types (i.e. 6, 11, 16, and 18) plus five additional cancer-causing types (i.e. 31, 33, 45, 52, and 58)[36]. Cervarix includes only the two most important high-risk types, HPV 16 and 18[37].

According to the main guidelines, pre-adolescent girls (9–15 years) can receive a two-dose HPV vaccine at either a six-month or one-year series interval, acquiring protection from HPV16, the most common type associated with cervical cancer, and other less common types. Vaccination protects from HPV infection until girls enter the routine screening program (HPV testing and cytology). Since 2015, the World Health Organization (WHO) recommends a two-dose program for younger girls, and a three-dose program for women of 15 years and older.

The primary target of vaccination is pre-adolescent girls, ideally before sexual debut. Other potential targets for the vaccination are females >15 years old, males and other high-risk individuals (e.g., HIV-positive patients), whether affordable and sustainable. However, the main goal is to achieve high vaccination coverage in pre-adolescent girls[38].

In the last ten years, Gardasil was used and showed to be effective in preventing HPV 16 and 18 infection in countries where high coverage was reached. After, Gardasil9 has replaced Gardasil. Gardasil9 has the same rapid anti-HPV 18 and HPV45 titer loss as Gardasil did. Cervarix, includes HPV16 and 18 L1 and has different adjuvants than Gardasil. It showed to be effective in the prevention of HPV infections as well, maintaining high antibody titers for at least ten years. Even just one dose of Cervarix protects against HPV 16 and 18 infections with robust antibody titers, offering a cost-effective vaccination program, especially in developing countries.

	Gardasil	Gardasil9	Cervarix
Oncogenic protein subunit component L1 VLP, μg			
HPV 16	40	60	20
HPV 18	20	40	20
HPV 31		20	
HPV 33		20	
HPV 45		20	
HPV 52		20	
HPV 58		20	
Verrucous protein subunit component L1 VLP, μg			
HPV 6	20	30	
HPV 11	40	40	
Manufacturing components			
Sodium chloride, mg	9.56	9.56	4.4
L-Histidine, mg	0.78	0.78	
Polysorbate 80, μg	50	50	
Sodium borate, μg	35	35	
Sodium dihydrogen phosphate dihydrate, mg			0.624
Adjuvant			
Amorphous aluminum hydroxyphosphate sulfate, μg	225	500	
3-*O*-Desacyl-4′-monophosphoryl lipid (MPL) A, μg, adsorbed on			50
Aluminum hydroxide salt, μg			500
Expression system			
Recombinant *Saccharomyces cerevisiae*	Yeast	Yeast	
Trichoplusia ni insect cells			Baculovirus

Table 3: HPV vaccines composition[39,40]

Screenings and vaccination programs have an essential role in reducing the incidence and mortality of cancers related to HPV infection[41–44].

In the past ten years, most developed countries, e.g., many European countries, have conducted HPV vaccination programs targeting young girls, with coverage between 30 and 80%[45].

Australia represents an excellent example as vaccination coverage is above 80% for girls and 75% for boys. Cervical cancer is expected to be almost eradicated by 2066, assuming the maintenance of this vaccination coverage and a 5-yearly HPV testing for cervical cancer screening[46].

In the United States, the human papillomavirus (HPV) vaccine was introduced for females 11-12 years old in 2006 and also recommended through age 26 years for women not vaccinated previously.

In the beginning, three doses were recommended in the USA. Now just two doses are considered sufficient for girls younger than 15 years, while three doses are kept for older women.

In 2013, a median of 12% and 19% of adolescent girls covered by commercial health plans and Medicaid plans, respectively, had completed the vaccination series by age 13[47].

In 2011, the United States was the first country to adopt a gender-neutral routine HPV immunization policy. Thus, routine HPV vaccination was also recommended for boys aged 11 or 12 years and for those through 21 years not vaccinated previously[48].

In France, HPV vaccination is promoted especially in girls aged 14 with a catch-up program for females from 15 to 23 years old. In this country, there is a reimbursement of the 65% of the vaccine price, resulting in Gardasil® being the fifth-highest drug expenditure of the main scheme of the French National Health Insurance in 2008. The maximum uptake in the catch-up group, for both 1 and 3 doses, was observed for women born in 1992 (15 years in 2007) with 52.5% and 35.6%, respectively[49].

In Italy, since 2008, vaccination is offered for free to all young girls aged 11 years. In 2017, with the release of the National Vaccine Prevention Plan (PNPV) 2017-2019, vaccination was extended to boys in the twelfth year of life, men who have sex with men (MSMs), and

immunocompromised patients (e.g., HIV-positive patients)[46]. These new recommendations are aligned with the recent evidence about the cost-effectiveness of HPV vaccination in other targets[50].

The target set by the PNPV is to achieve a 95% vaccination coverage. However, in Italy, the latest full-course vaccination coverage was 50% among females and 20% among males[51].

The introduction of vaccination programs in less developed countries is harder to achieve. However, good results have been obtained in countries like Rwanda and Bhutan[18]. The promotion of HPV vaccination programs in less developed countries is of paramount importance, considering that in these countries, HPV infection is the leading cause of at least 50% of the infection-related cancers in both sexes[18].

The high incidence of HPV and HIV infections in less developed countries, associated with limited screening and vaccination programs, increases the rates of cervical cancer[52].

The socioeconomic status of a country has a significant impact on the infection-associated cancers burden, and specific population-based measures are required to reduce the incidence of these cancers worldwide[18].

However, compared to other carcinogenic infections, HPV is less linkable with socioeconomic development even if it remains a leading cause of infection-related cancers, especially in less developed countries[52].

.

3 Human Papillomavirus

Human papillomaviruses, a member of the *Papillomaviridae* family, are small viruses with a tropism for mucosal or cutaneous squamous epithelia. Some of these viruses are associated with the development of human cancers in different anatomical areas, such as the anogenital tract, and head and neck[53,54].

These viruses possess a non-enveloped icosahedral capsid (Figure 3) and a double-stranded circular DNA genome, which ranges in length from 5,748 bp for *Sparus aurata* papillomavirus 1 (SaPV1) to 8,607 bp for canine papillomavirus type 1 (CPV1)[55].

3.1 Genomic organization and viral proteins

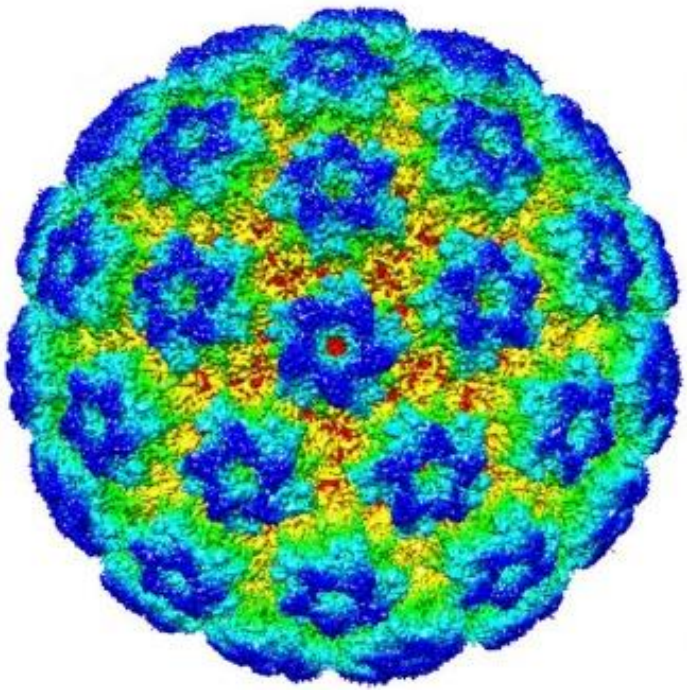


Figure 3: Human papillomavirus type 16 structure. Three-dimensional reconstruction[56]

The genome is organized in three major regions: the early, the late, and the long control region (LCR) that are separated by two polyadenylated (pA) sites called early pA (AE) and late pA (AL) sites[57].

The early region contains up to seven ORFs encoding viral regulatory proteins (E1, E2, E4, E5, E6, E7, and E8), and the late region encodes the two viral capsid proteins (L1 and L2) (Figure 4)[57,58]. The LCR contains the origin of DNA replication and transcription control sequences[58].

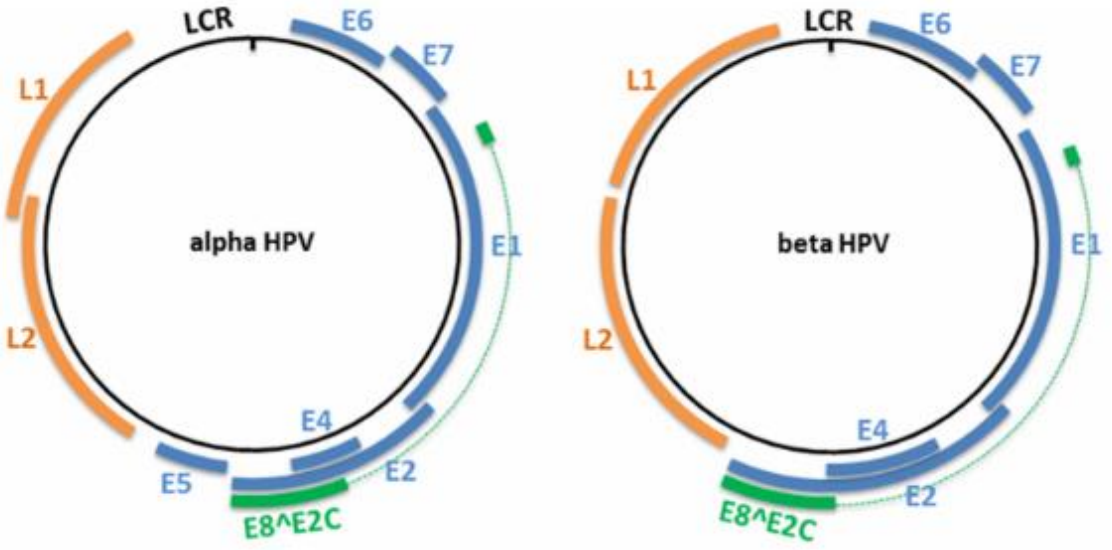


Figure 4: Organization of the viral genome: alpha (left) and beta (right) HPV types[55]

Only four ORFs (those of E1, E2, L1, and L2) are essential to fulfil the requirements ensuring the viral replication and shedding of the virus and are present in all known PVs[59].

However, some HPVs from different genera lack an ORF. For example, the E5 is lacking in beta, gamma, and mu genera, while both E5 and E6 are missing in HPV101, 103, and 108 from genus gamma[60,61]. The alpha HPV6 and 11 encode for E5γ and E5δ, two E5-like proteins[59].

HPV types belonging to beta and gamma genera present a shorter LCR compared to alpha HPVs (Figure 4, right).

All PVs have the potential to express a particular transcript, E8∧E2C, that encodes for a fusion protein composed of E8 fused to the hinge and DNA-binding domains of E2 (Figure 4)[62].

3.2 HPV lifecycle and natural history of the infection

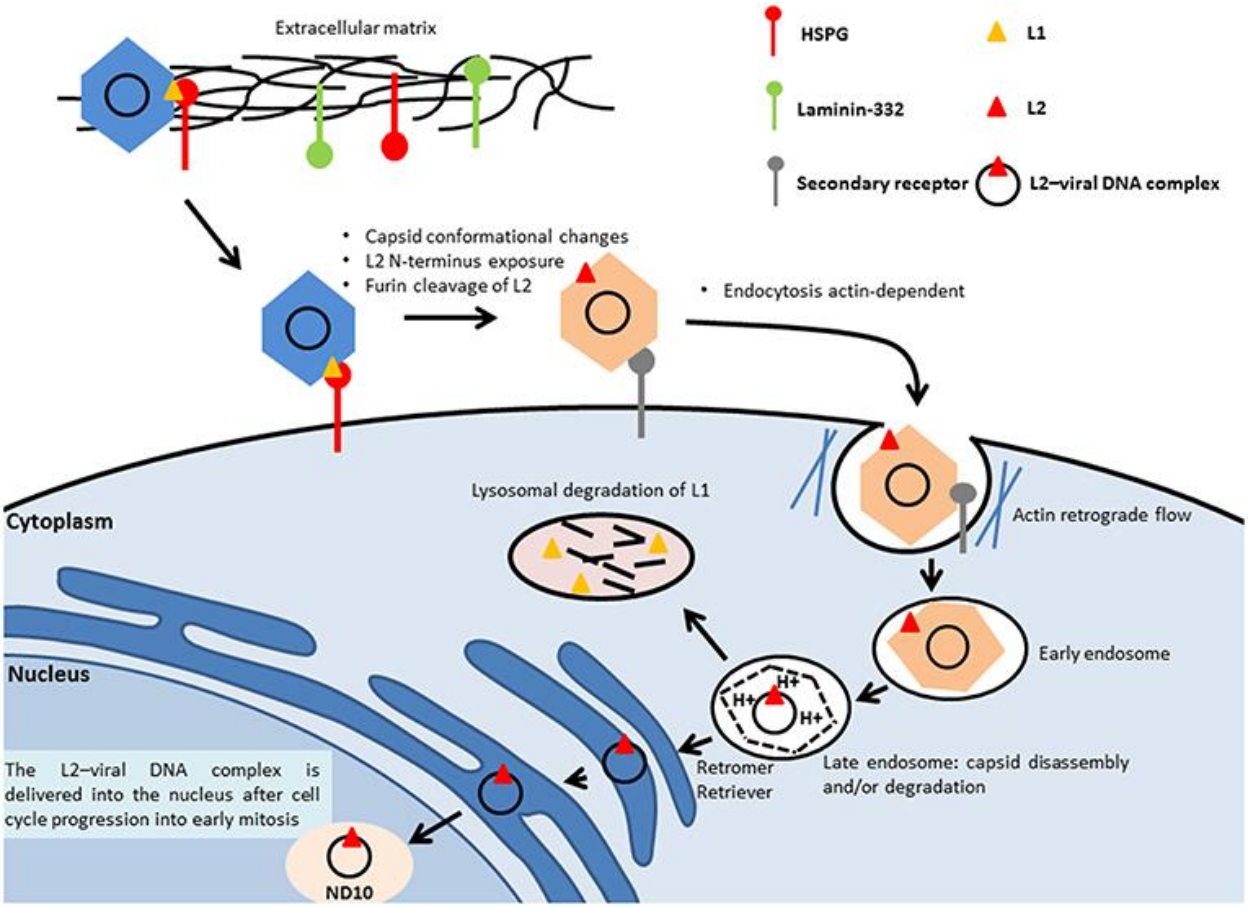


Figure 5: Early phases of the HPV infection. First, the virus enters in contact with the host cell membrane. After the internalization of the viral particle, the viral DNA is delivered into the nucleus of the infected host cell[55]

The virus has to reach the basal layer of the epithelium, to infect the basal keratinocytes or the epithelial stem cells. This is achieved by passing through micro-wounds or micro-fissures, or via hair follicles[63–65].

The squamocolumnar junction is the place where endocervix and ectocervix merge. In this portion of the cervix, the columnar epithelium flanks the squamous non-keratinized epithelium[66].

In prepubertal girls, the squamocolumnar junction is placed in the cervical canal. After puberty and during pregnancy, it extends outward over the ectocervix, and the junction gets exposed to the acidic vaginal environment[67]. In this context, the columnar epithelium can undergo physiological metaplasia and turns into the metaplastic squamous epithelium. This process causes the moving of the squamocolumnar junction in the inner part of the cervical canal, and all this metaplastic area is called transformation zone[68,69]. The metaplastic epithelium is more susceptible to HPV infection, thus prepubertal and mainly pubertal girls are considered at higher risk of contracting HPV infection. The sexual contact is considered the primary way to contract HPV infection.

HPV infection has a higher prevalence in girls younger than age 20, and after this age, there is a significant decrease in the number of positive individuals[70,71].

In the oral cavity, the tonsillar crypts have a similar cellular structure to the transformation zone of the cervix, facilitating the HPV infection[72].

In a recent publication, Broniarczyk and colleagues showed that the viral particles could stay infectious, even after several weeks of permanence on the surface of senescent cells, resistant to the viral infection. This work showed that the reactivation of the cell cycle, mediated by p53 siRNA, allowed the entry of the virus[73].

The virus, with the major capsid protein L1, binds the heparin sulfate chains of proteoglycans (HSPGs), placed in the cell membrane or at the extracellular matrix (ECM), and thus initiates the infection.

This interaction causes conformational changes of the capsid structure mediated by cyclophilin (CyP) B that exposes the minor capsid protein L2 on the surface of the virus (Figure 5).

A key event in this process is the cleavage of the L1 protein performed by kallikrein-8 serine protease[74].

Laminin-322 on the ECM can allow a transient binding of the virus at the early stages of the infection[75].

After, a conserved furin consensus site of the L2 protein is cleaved, allowing the exposure of the RG1 neutralizing epitope (Amino acids 17 to 36). This step seems to be important for the interaction of the virus with an unidentified secondary receptor, and thus for the internalization of the virus[76,77].

Actin protrusions on the cell membrane showed to play a role in the transport of the HPV16 virions inside the cells by actin retrograde flow[78,79]. Furthermore, actin polymerization and depolymerization showed to be significant events in the HPV16 endocytosis process, in particular for scission of endocytic vesicles[79].

Once inside the cell, the virus is delivered to an early endosome, that matures into a late endosome[80] where the capsid is disassembled in a low pH environment by host-cell CyPs, leading to the dissociation of L1 and L2 proteins[81]. Based on a recent study, residual L1 proteins can remain in complex with the viral DNA[82].

The L2 protein mediates the endosome exit of the viral genome that is transferred to the trans-Golgi network, by the retromer complex[83].

The Retromer complex plays a role in the transfer of different cargo proteins from the endosome to the trans-Golgi network. It is composed of sorting nexins-dimers (SNX1, SNX2, SNX5, and SNX6) and a vacuolar protein sorting (Vps) trimer containing Vps26, Vps29, and Vps35. The retriever complex and the cellular adaptor protein SNX17 are also involved in this process of endosomal trafficking[84]. A study on the SNX17 showed that this protein interacts with HPV16 L2[85]. SNX17 binding site is conserved among different PVs, thus may play a key role in regulating the PV life cycle and replication.

The interaction between SNX17 and the viral L2 proteins is essential to protect viral capsids from lysosomal degradation and also for an effective capsid disassembly[86].

SNX27, part of the retromer complex, interacts with L2 through its PDZ domain and is important for the virion trafficking process[87].

L2 plays an essential role in the trafficking of the viral genome to the nucleus[83]. The L2 in complex with the viral DNA is translocated into the nucleus[88] together with residual L1[55].

The transfer of the L2-viral DNA complex into the nucleus requires a cell in early mitosis phase[89], with a disrupted nuclear envelope[90] (Figure 5).

Once inside the nucleus, the viral genome is delivered to the ND10 domain (the promyelocytic leukemia bodies), and the viral transcription and replication can start[91,92].

ND10 domains act as a natural defense against viral infections[93] repressing transcription, replication, and establishment of incoming HPV DNA in the early stages of infection[94].

L2 indices the alteration of the ND10 protein composition, and this causes the release or degradation of Sp100[95].

The Daxx protein is part of the ND10 complex and seems to be involved in the early gene expression and the transient replication of the viral genome[96].

The replication of the virus is dependent on the differentiation status of the epithelium. The first step of this process is the establishment of the replication, where the maintenance of a constant number of episomal copies is established (50-100 per cell) (Figure 6)[97].

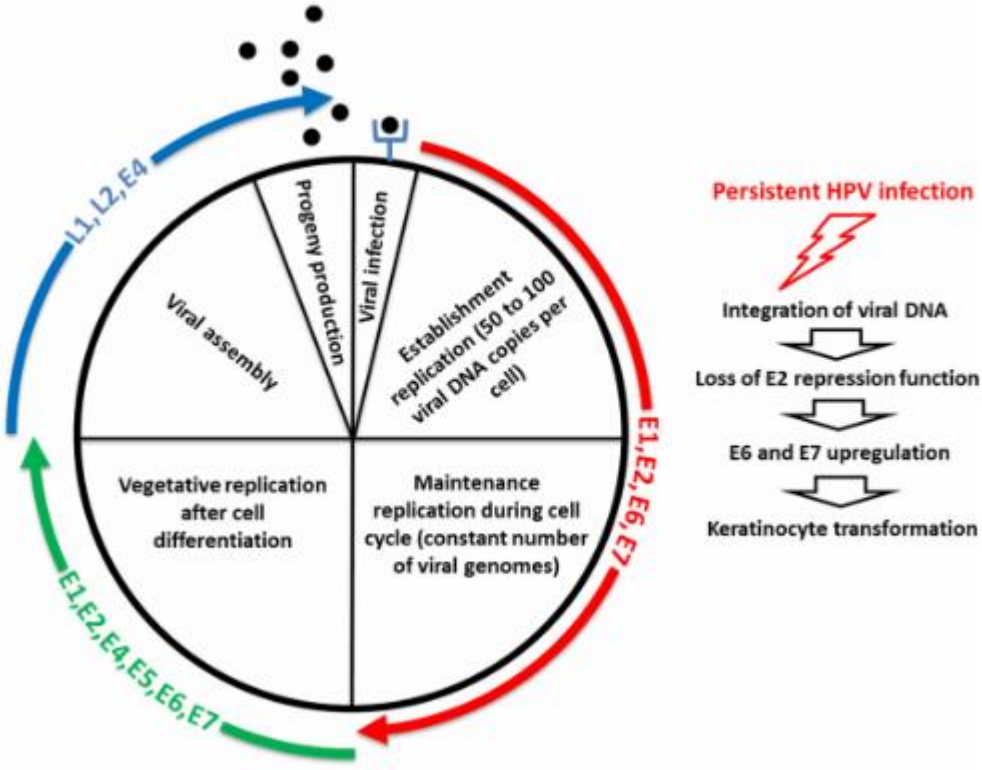


Figure 6: Schematic representation of the main phases of the HPV life cycle[97].

E1 and E2 early proteins support the viral DNA replication, which relies on the host DNA replication machinery[98].

Once the viral DNA is transferred into the nucleus, E2 binds specific sites of the LCR, recruiting E1 to the origin of replication, and the viral DNA replication is initiated[99]. Cell proteins like TopBP1 and Brd4 are involved in the HPV16 DNA replication process[100].

The interaction of NCOR/SMRT repressor complexes with E8∧E2C proteins inhibits the viral replication, and limited viral genome amplification is obtained[101,102]. After this initial step in which the viral genome is replicated in low copies, the maintenance phase starts (Figure 6)[97].

In this phase, in the nucleus of undifferentiated basal cells, the viral genome in extrachromosomal form is replicated with the perspective of establishing persistent infection.

The attachment of the viral genome to cellular chromosomes helps the correct segregation of the viral genome during the cell division.

E2, through the transactivation domain at the N-terminus, and Brd4 protein, through the bromodomain, both interact with the mitotic cell chromosomes. Brd4 is known to interact with the E2 of most PVs[103].

Simultaneously, the DNA binding/dimerization domain of E2 binds the LCR of the viral genome at the E2 binding sites[104–108].

The ATP dependent DNA helicase ChIR1 seems to be involved in the interactions between the chromatin and HPV16 E2 and, also, in the maintenance of the episomal form of the viral genome[109–111].

The SMC5/6 complex seems to influence the maintenance of viral episomal DNA by interacting with the papillomavirus E2 protein[112].

Also, E6 and E7 oncoproteins play a role in the maintenance of the viral episomal form[113,114].

Following the cell differentiation in the stratified epithelium, after the initiation of the viral genome replication, progeny virions are produced[97,115,116] (Figure 6).

In normal conditions, in a non-infected epithelium, cells leave the basal layer for terminal differentiation, exit the cell cycle, and stop the replication of the DNA.

Papillomavirus has developed strategies to prevent the interruption of the cell cycle, and to inhibit apoptosis signals, favouring the conditions necessary for the viral DNA replication and the production of viral particles.

The inactivation of tumor suppressor proteins (e.g., p53, pRb) and the induction of other activating signals controlled by HPV E6 and E7 that are expressed at relatively low levels in differentiated cells induce cells activation and progression in S phase.

In this context, the late promoter (P670 for HPV16, P811 for HPV18, and P742 for HPV31), located in the E7 region, is activated, inducing the expression of high levels of E1 and E2 viral proteins required to ensure viral DNA replication. Also, the activity of E4 and E5 is essential for an efficient, productive replication[117,118].

The expression of L1 and L2 capsid proteins is controlled by the late promoter. Their expression leads to the packaging of the newly replicated viral DNA. Finally, virions are released in the superficial layer of the epithelium during desquamation[119] (Figure 6).

The interaction of E4 with the keratin network is also crucial in this part of the viral life cycle[120].

Furthermore, epigenetic mechanisms regulate access to viral genes, controlling their espression[121]. CTCF (CCCTC-binding factor) and YY1 (Yin Yang 1) are critical factors in this process. CTCF binds to E2 ORF and YY1 to the viral LCR.

CTCF and YY1 factors are involved in the formation of a loop and, therefore, in the interaction of E2 with the LCR. This event induces epigenetic repression of the chromatin in HPV18, resulting in attenuated expression of oncoproteins in undifferentiated cells.

During cell differentiation, a drop in YY1 expression levels leads to loss of chromatin loop formation, inducing the expression of oncoproteins[121].

3.3 Transforming activities of the oncogenic papillomaviruses

The study of high-risk HPV types (mainly HPV16 and 18) allowed the identification of the main factors inducing cervical cancer development. E6 and E7 oncoproteins, from oncogenic HPV types, were identified as the main factors involved in this process. These oncoproteins act by altering pathways involved in host immune response and cellular transformation.

E6 and E7 oncoproteins alter the regulation of the cell cycle and apoptosis control by interacting with many cellular proteins.

Also, E5, a small hydrophobic oncoprotein, seems to play a role in the malignant progression.

E6 and E7 need to be constitutively expressed to induce cellular transformation. Experiments conducted to evaluate the effect of E6 and E7 inhibition in HPV-positive cancer cells resulted in cell growth arrest and induction of apoptosis or senescence[122,123].

E6 acts inhibiting the activity of p53, while E7 inhibits pRb. p53 is a key factor in the DNA damage response and apoptosis, while pRb regulates cell cycle control. While the activity of E6 and E7 from HR HPV types was extensively studied, less clear is the role of the corresponding proteins from beta HPV types in human cancers. A different mechanism seems to control skin cancer development since cutaneous E6 and E7 transcripts are not detected in skin tumors. A hit-and-run mechanism is hypothesized to explain this process.

Moreover, E5 is not present in the genome of beta and gamma types, confirming the idea of a different mechanism of tumor induction[55].

Cancers associated with high-risk HPV types infection often show the expression of viral proteins without the production of new viral particles. These non-productive infections could occur after the integration of viral genomes into the DNA of infected host cells. This integration can disrupt

the E2 gene that has a role in the negative regulation of E6 and E7 expression[124,125]. Thus, after this event, the expression of E6 and E7 is upregulated.

In many cervical cancers, HPV16 genomes are found integrated into the host cell chromosome. A study demonstrated that integration of HPV16 DNA leads to increased steady-state levels of mRNAs encoding the viral oncogenes E6 and E7. The integration of viral genomes into the host chromosome disrupts the 3' UTR of the early region of the viral genome, increasing the stability of E6 and E7 mRNAs. Furthermore, A+U-rich element within the 3' UTR seems to contribute to the instability of the heterologous mRNA.

Thus, integration of HPV16 DNA can result in the increased expression of the viral E6 and E7 oncogenes through altered mRNA stability, inducing cervical cancers development[126].

Genome-wide analyses of cancers related to HPV infections showed integration sites often flanked by chromosomal aberrations, including focal amplifications, rearrangements, deletions, and translocations[127]. These alterations can be the result of the genomic instability induced by the activity of E1 and E2. E1 and E2 are capable of causing over-amplification of the genomic locus where HPV origin was integrated. Genomic analyses of cells with integrated HPV genomes showed excision, rearrangement, and *de novo* integration of the HPV and flanking cellular sequences. The papillomavirus replication machinery, with this mechanism, can induce genomic changes of the host cell that may facilitate cancinogenesis[128,129]. A "looping model" describes how viral-host DNA concatemers can be generated after the formation of a loop that acts as a substrate for rolling circle replication[127]. Other rearrangements could be due to the intrinsic genomic instability of a specific host region and from other E6/E7-mediated genetic instability[130].

If the integration of the viral genome into the host cell chromosome seems to have a significant impact on the ability of the virus to induce cancer development, in some cervical cancers, the viral genome is maintained in episomal form. In this context, the deregulation of viral genes seems to

occur through another mechanism, which implies the accumulation of aberrant epigenetic mutations on the viral genome[131–134].

Several studies highlight that overexpression of high-risk HPV E6 and E7 oncoproteins induces the progression of the neoplastic process[57,135].

In healthy human cells, spontaneous mutagenesis occurs with low frequency, but the expression of high-risk HPV E6 and E7 oncoproteins triggers the genomic instability[136].

The expression of the high-risk HPV E6/E7 genes first induces the accumulation of premalignant alterations and, at a later time, directly contributes to malignant progression by promoting genomic instability[137,138].

Finally, in cervical carcinomas, mutations in recurrent host cell genes have been identified. These include EP300, FBXW7, PIK3CA, HLA-B, TP53, MAPK1, PTEN, ERBB2, NFE2L2, and STK11[139].

3.3.1 Structure and function of E6

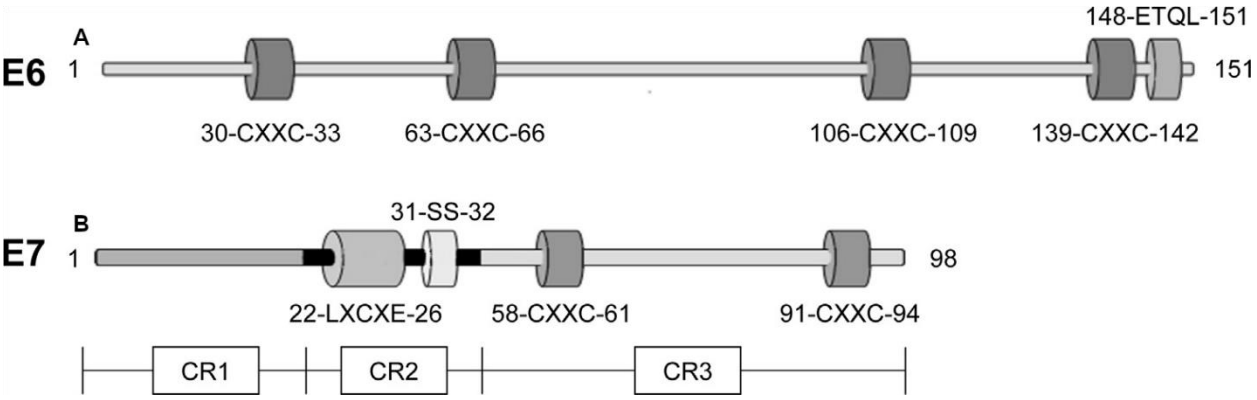


Figure 7: High-risk HPV E6 and E7. Essential amino acid motifs for the function of the E6 protein are shown. Both proteins contain zinc fingers (CXXC motifs) domains that are important for the interaction with host cell proteins. E6 (A) has a consensus PDZ-binding motif (ETQL) at the C terminus. E7 (B) is divided into three domains: CR1, CR2, and CR3. Regions involved in pRb binding (LXCXE) and the two serine residues (Ser31 and Ser32) that are susceptible to casein kinase II (CKII) phosphorylation are shown[54].

E6 is a protein of approximately 18 kDa with two zinc finger domains composed of two CXXC motifs each, essential for the protein function (Figure 7A)[140,141].

High-risk HPVs E6 interacts with the ubiquitin ligase E6AP inducing degradation of p53 through the proteasome pathway. This interaction causes a conformational change that allows the formation of the complex E6/E6AP/p53, leading to the rapid degradation of p53[142]. E6 can also interact with CBP and p300 transcriptional co-activators, inhibiting the expression of p53-regulated genes[143] (Figure 8). In mucosal low-risk, high-risk, but also in some cutaneous types, including some beta types, the apoptotic response is inhibited by inducing degradation of Bak, a member of the Bcl-2 family[144–146] (Figure 8). The interaction with E6AP and thus, the induction of the proteasome pathway contributes to this process. Studies on the skin cell lines showed that Bak is stabilized and induced in response to UV irradiation. Thus, the Bak inhibition results to be a pivotal event to contrast the anti-proliferative effect of the UV irradiation[54] (Figure 8). Members of the guanylate kinase MAGUK family are also targets of the high-risk HPV E6[147].

Survivin, a member of the inhibitor of apoptosis (IAP) gene family, suppresses apoptosis and controls cell division. HPV 16 E6 and E7 oncoproteins showed to upregulate endogenous survivin mRNA[148].

HPV 16 E6 oncoprotein also binds to tumor necrosis factor TNF R1 preventing cells from undergoing TNF-induced apoptosis[149]. HPV 16 E6 protein also binds to the Fas-associated death domain (FADD), protecting cells from apoptosis[150]. Furthermore, HPV 16 E6 oncoprotein seems to promote the activity of caspase 8 in the nucleus. Caspase activity appears to be essential for the later stages of the viral life cycle in differentiating keratinocytes[151,152].

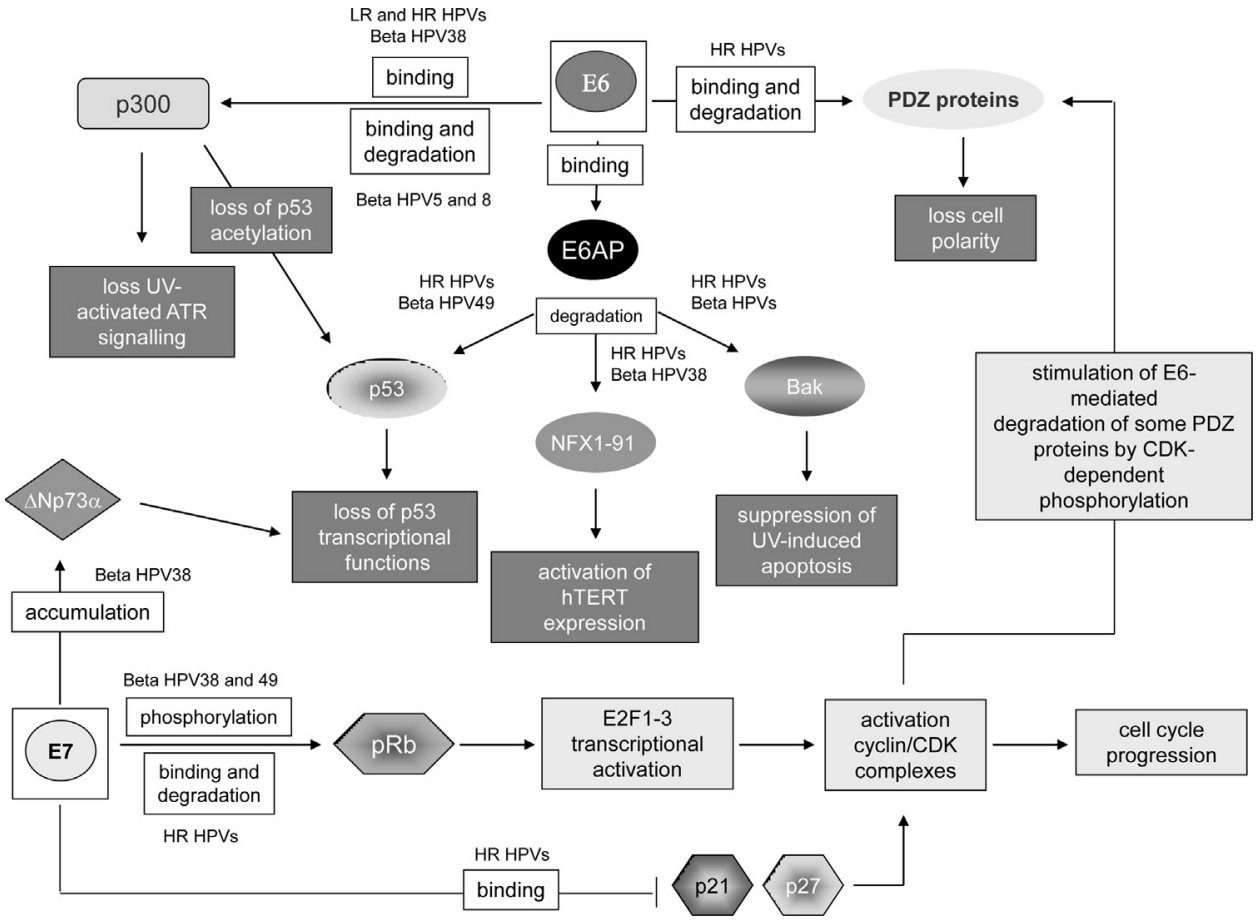


Figure 8: E6 and E7 interactions and activities. In dark grey events induced by the E6, while in light grey the ones mediated by E7[54]

3.3.2 Structure and function of E7

E7 is a protein of around 12 kDa with a zinc-binding domain at the C-terminus, important for the structural integrity of the protein[153,154] (Figure 7 B).

A typical E7 protein owns three principal domains: CR1, CR2, and CR3 (Figure 7 B). CR1 provides the ability to induce S-phase progression and cellular transformation[153,155]. CR2, through the LXCXE domain, mediates the interactions with pRb protein and its related proteins, p107 and p130. These proteins control the cell cycle. E2F1-3, a member of the E2F family, is inhibited by pRB1, thus maintaining the cell in a quiescent state during the G0/G1 phase (Figure 8).

Association of high-risk HPV E7 with pRb leads to degradation of pRb via proteasomal pathways and transcription of E2F, leading to cell cycle progression (Figure 8)[156].

In many HPV types, the E7 CR2 domain holds two serine residues (Ser31 and Ser32) susceptible to casein kinase II (CKII) phosphorylation[54] (Figure 7 B).

CR3, which contains the two CXXC motifs, is involved in the interaction with different proteins such as p21WAF1/CIP1 and p27KIP1, causing neutralization of their inhibitory effects on the cell cycle[157–159] (Figure 8). E7 interacts with several other transcription factors like AP-1, TBP, or c-Jun, inducing the disruption of cell cycle control[160].

3.4 Classification of the Papillomaviruses

The classification of the Papillomaviridae family is based on pairwise nucleotide sequence identity across the L1 open reading frame, as it is reasonably well conserved in all known PVs, allowing a genome-based approach to PV nomenclature (Figure 9).

Moreover, PVs do not induce robust antibody responses, and thus neither a “serotype-based” classification can be used. In consequence, the classification of PV types is based only on nucleotide sequence similarities[161–165].

In the 7th report of the International Committee on Taxonomy of Viruses (ICTV) held in 2002, Papillomaviruses was designed as a distinct family of viruses[166]. The work of de Villiers et al. (2004)[165], proposed a classification of 92 human papillomavirus (HPV) and 24 animal PV types based on the guidelines established by the ICTV and the PV research community and this publication consolidated the classification method that was formalized later, in the 8th Report of the ICTV in 2005[167].

According to these guidelines, *Papillomaviridae* members are divided into genera designed by Greek letters and into species by addition of a number to the letter.

Different genera share less than 60% nucleotide sequence identity in the L1 ORF. Species within a genus share between 60% and 70% nucleotide identity. PV types within a species share between 71% and 89% nucleotide identity within the complete L1 ORF[165,168].

Moreover, PVs sharing between 90% and 98% nucleotide identity within the complete L1 ORF are named subtypes, while those sharing >98% L1 ORF nucleotide identity are named variants[168] (Figure 9).

In 2010 Bernard and colleagues performed an alignment analysis comparing the L1 genomic sequences of 189 HPV types and PVs from non-human mammals, birds, and reptiles (64, 3, and 2 PV types, respectively)[168]. They described a total of 29 genera (Figure 9), which confirmed the consistency of the current classification method based on the L1 genomic sequence.

The Papillomavirus Episteme (PaVE) is a database of curated papillomavirus genomic sequences.

According to the rules of the PaVe database, to be recognized as a novel papillomavirus type, a viral genome has to fulfil strict requirements[165,168]: (i) The entire viral genome must be cloned; (ii) the L1 sequence cannot share >90% nucleotide sequence identity with its closest neighbour and (iii) the cloned genome must be submitted to and reviewed by the International Human Papillomavirus Reference Centre[169].

If the cloned viral genome respects the parameters required, the International Human Papillomavirus (HPV) Reference Center assigns a HPV type number deposits. It maintains the reference clones, as well as distributes samples of the reference material for research use.

With the advent of the NGS, in the last few years, several novel HPVs have been described[170] that do not meet all these requirements and will therefore not be recognized as novel viral types by the International Human Papillomavirus Reference Centre.

According to the Human reference clones database of the International Human papillomavirus (HPV) Reference Center (https://www.hpvcenter.se/human_reference_clones/), the whole genome of 228 different HPV types was identified and cloned. This database includes 65 Alphapapillomaviruses, 54 Betapapillomaviruses, 99 Gammapapillomaviruses, 3 Mupapillomaviruses, and 1 Nupapillomavirus (update April 2020)[171].

Instead, according to the PaVe database (https://pave.niaid.nih.gov/), regarding the HPV types that were discovered and fully characterized, but not necessarily cloned into a vector, there are 66 Alphapapillomaviruses, 67 Betapapillomaviruses, 301 Gammapapillomaviruses, 5 Mupapillomaviruses and 1 Nupapillomavirus, discovered so far (update May 2020).

According to the PaVe database, updated on May 2020, genus alpha includes 14 species (plus an additional group of unclassified alpha types), genus beta consists of six species (plus another group of unclassified beta types), genus gamma contains 27 species (plus an additional group of unclassified gamma types), genus mu includes three species and genus nu includes one species. Additionally, more than 200 “not referenced” HPV genomes are present in the PaVe database (update May 2020); thus, the numbers of species belonging to the different genera will probably grow[171,172].

Moreover, the NCBI Nucleotide database, a collection of sequences from several sources (e.g., *GenBank*, RefSeq, TPA, and PDB)[173,174], abound of “partial HPV sequences” (roughly 20,000 in May 2020), representative of many other human papillomaviruses that need to be fully characterized.

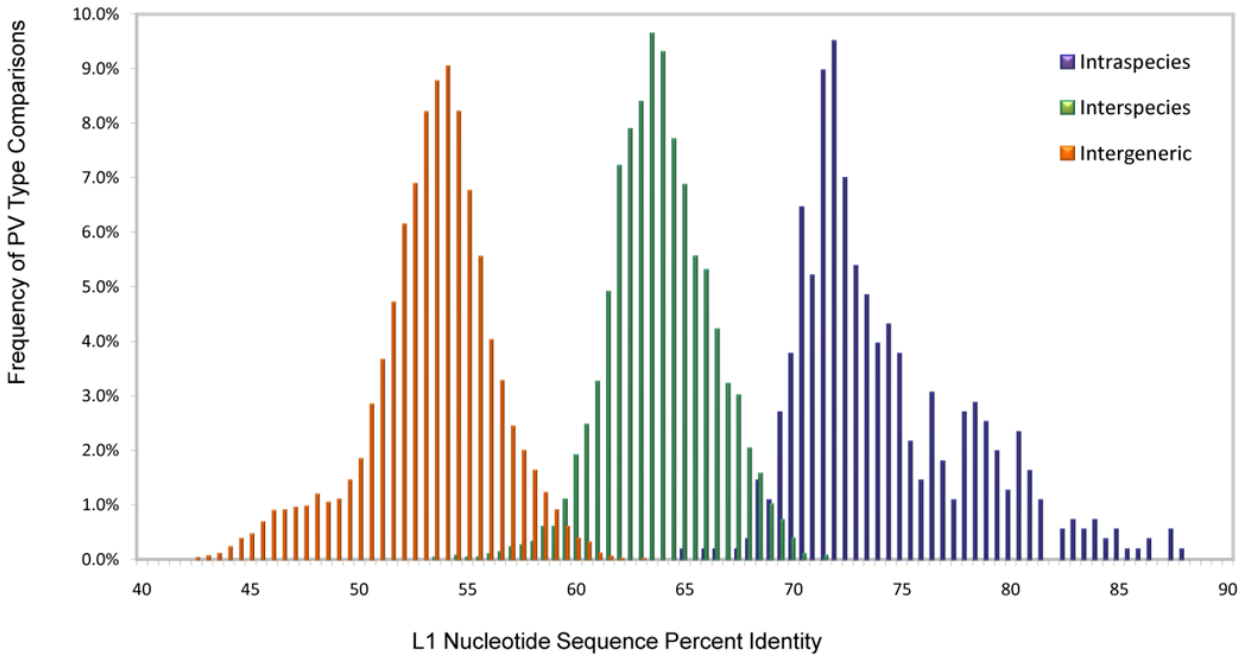


Figure 9: A global multiple sequence alignment matrix was generated by aligning the nucleotide L1 sequences of 189 different HPV types. L1 nucleotide global multiple sequence alignments were guided by amino acid alignments. Different clusters of identity percentages can be appreciated, according to intergeneric, interspecies, or intraspecies comparisons. In the X-axes the L1 percentage of identity and Y-axes the percentage of the total number of comparisons[168]

3.5 Tropism of the Human papillomaviruses

Epithelial tissue is composed of three elements: epidermis, dermis, and subepidermal tissue (subcutaneous tissue at skin sites). The epidermis is composed of multiple keratinocyte layers, and many papillomaviruses are detected in this epithelium. Moreover, the skin presents other functional appendages, including hair follicles, sebaceous glands, eccrine and apocrine sweat glands, nails, and the inter-appendageal epidermis. Specialized epithelial sites contain additional appendages, such as the salivary glands of the oral cavity and the tonsillar crypts of the oropharynx. For papillomaviruses, these “specialist” structures represent particularly vulnerable sites, as they lack the highly structured barrier function usually associated with the epithelium.

The transformation zones have an even more complicated structure. Here, squamous epithelium and columnar epithelium intersect, and these areas are particularly susceptible to cancer-related HPV infections[175].

Nowadays, alpha, beta, and gamma are the most represented HPV genera and are subdivided into two categories according to their tropism: mucosal and cutaneous types.

In table 4 are reported the main HPV types, their tropisms, and the associated diseases[53].

Approximately 40 alpha HPVs infect mucosal epithelia and are divided into two groups, low-risk, and high-risk types, based on their association with benign warts or lesions that have a propensity for malignant progression[53].

Genus	Species	Representative HPV types	Tropism	Associated Diseases
Alpha-PV	α**1**	32	mucosal	Heck's disease
	α**2**	3, 10, 28	cutaneous	flat warts
	α**4**	2, 27, 57	cutaneous	common warts
	α**7**	18, 39, 45, 59, 68	mucosal	intraepithelial neoplasia, invasive carcinoma
	α**9**	16, 31, 33, 35, 52, 58	mucosal	intraepithelial neoplasia, invasive carcinoma
	α**10**	6, 11	mucosal	condyloma acuminate
		13		Heck's disease
Beta-PV	β**1c**	5, 8, 12, 14, 19, 20, 21, 24,25, 36, 47	cutaneous	Epidermodysplasia verruciformis
	β**2**	9, 15, 17, 22, 23, 37, 38	cutaneous	Epidermodysplasia verruciformis
	β**3**	49	cutaneous	Epidermodysplasia verruciformis
Gamma-PV	γ**1**	4, 65	cutaneous	Warts
	γ**4**	60	cutaneous	Warts
Mu-PV	**μ1**	1	cutaneous	plantar warts
	μ2	63	cutaneous	Warts
Nu-PV	**ν**	41	cutaneous	Warts

Table 4: Summary of the main HPV genotypes, their tropism, and related diseases[53]

A subgroup of 12 alpha types (IARC Group 1: HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59), based on epidemiological and biological data, have been classified as high-risk carcinogenic types. Additional eight types (IARC Groups 2A and 2B: HPV26, 53, 66, 67, 68, 70, 73, and 82), have been classified as probably or possibly carcinogenic[176,177].

Furthermore, the alpha genus also includes a few cutaneous HPV types (HPV2, 3, 7, 10, 27, 28, and 57), which cause common and plantar warts[178–181].

Cutaneous HPVs are dispersed across all five HPV genera and have most frequently been detected in healthy skin samples, suggesting their commensal nature[170,182].

The majority of the HPV types belonging to genus beta have a cutaneous tropism. Around 50 different beta HPV types, divided into five species, have been fully characterized so far. They are primarily found on the surface of the skin[183].

Nevertheless, many studies detected beta HPV types at different anatomical sites other than the skin, such as the oral mucosal epithelium, genital sites, and the anal canal[184–187].

In particular, beta 3 HPV types were identified both in cutaneous tissues and mucosal epithelia, suggesting a dual tropism[186,188].

Moreover, different studies showed that beta 3 HPV types share biological similarities with mucosal HR HPV types such as HPV16 *in vitro* and *in vivo* experimental models[189,190].

Gamma types are known to have a cutaneous tropism, but the growing number of types belonging to this genus, highlight the need to investigate more on the tropism and biology of these viruses[55,191]. Gamma 6 appears to be only in mucosal epithelia[191].

Recent studies showed the presence of gamma types in additional anatomical sites, other than the healthy skin, including cutaneous and mucosal lesions and healthy mucosa, suggesting their

double, mucocutaneous tissue tropism, and adding more questions about their clinical importance[60,188,192].

Figure 10 shows a maximum likelihood phylogenetic tree obtained in 2012 by aligning sequences from 139 HPV types. The sequences were aligned at the amino acid level using the MUSCLE alignment tool considering only the E1, E2, L2, and L1 coding regions concatenated. RAxML with GTR + G4 model was used to generate the tree[97].

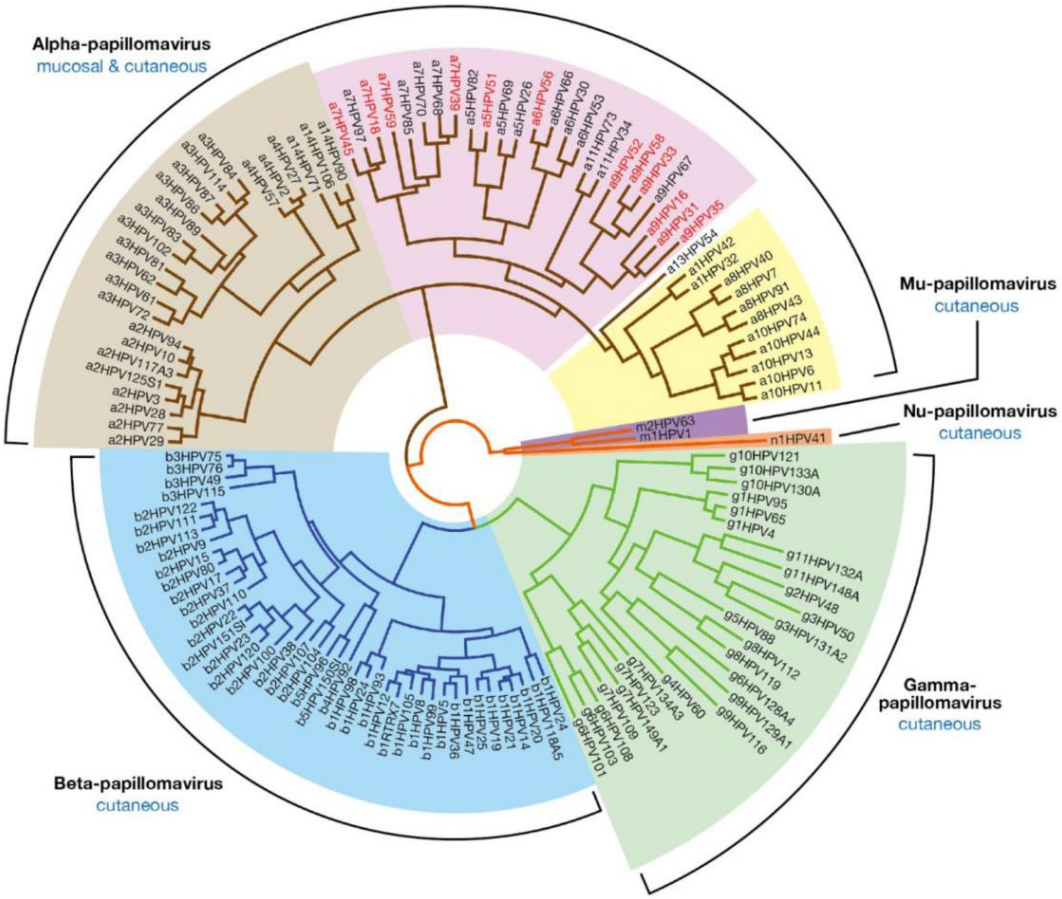


Figure 10: Evolutionary relationship between 139 HPV types. Only E1, E2, L1 and L2 ORFs have been considered for this analysis[97]

4 HPVs and human diseases

HPV is the most common infection sexually transmitted worldwide. The majority of sexually active individuals are infected with HPV at least once during their lifetime[193].

HPV infection is associated with a broad spectrum of pathologies from benign lesions (e.g., cutaneous warts, recurrent respiratory papillomatosis) to cancers[97]. One of the most critical risk factors for cancer development is the persistence of the HPV infection[97,193].

Long-lasting HPV infection is related to the occurrence of different kinds of cancers, particularly cervical, oropharyngeal, anal, penile, as well as vaginal and vulvar cancers. HPV16 and 18 are the most common HPV types detected in HPV related cancers[193,194]. Most of the people with a functional immune system can clear HPV infections with no consequences or recurrences[195–197]. Instead, people with a compromised immune system, like those HIV positive patients, AIDS, or patients undergoing solid organs transplantation, showed to be more susceptible to high-risk lesions development and subsequent progression to malignancy[195–197].

The most represented HPV genera detected in different anatomical areas of the human body are alpha, beta, and gamma[169,198].

HPV types belonging to the Alpha genus are of primary medical interest because of their clear involvement in different human mucosal cancers. Nowadays, the association of alpha HPV with vulva, vagina, penis, anus, and oropharynx cancers is widely accepted by the scientific community[193,199].

Alpha genus includes the low-risk types (LR) that are linked with the development of ubiquitous asymptomatic infections and benign papillomas and the high-risk types associated with the development of malignant lesions[200]. The high-risk carcinogenic types of HPV (Group 1) designated by the International Agency for Research on Cancer (IARC) in 2012 are HPV16, HPV18, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58, and HPV59.

HPV68 is classified as probably carcinogenic (Group 2A), and HPV26, HPV30, HPV34, HPV53, HPV66, HPV67, HPV 69, HPV70, HPV73, HPV82, HPV85, and HPV97 have been associated with rare cases of cervical cancer and are considered probable carcinogens (Group 2B)[201].

Beta HPV types are widely present on the surface of the skin of the general population and were firstly characterized in *Epidermodysplasia verruciformis* (EV) patients. In EV patients, recent studies indicate the involvement of cutaneous β-HPV types, together with UV radiation, in the development of skin cancer[55,202–207].

Gamma genus represents a large group of HPV types ubiquitously found on the skin of healthy individuals. Epidemiological and biological studies have been performed on the role of gamma HPV types, and do not support an etiological role in human skin cancers. Nevertheless, this genus includes an increasing number of members that may deserve more investigation. A member of species γ27 (HPV197) was isolated exclusively from skin cancers, using deep sequencing. However, additional studies are required to demonstrate an etiological link between gamma types and skin cancer[55,208,209].

4.1 Cutaneous and mucosal warts

Warts, also known as human viral papillomas, are benign epidermal tumors proliferations. Their infectious nature, due to human papillomaviruses (HPV), was first described by Jadassohn in 1896[210,211].

Seven different clinical types of warts are distinguished: Common warts (*verrucae vulgaris*), flat warts, deep palmoplantar warts, cystic warts, focal epithelial hyperplasia, Butcher's warts and anogenital warts (venereal warts, or condyloma acuminate)[212].

Common warts are found mainly on the dorsal side of hands and fingers, under and around nails, on the knees and the ankles; however, they may occur anywhere on the skin[212].

Common warts are associated with HPV types 2,4 (most common), followed by HPV types 1,3,27,29 and 57[213].

Flat warts have a higher incidence in children. They often occur in vast numbers and can even be spread out on the cheeks, the chin, the back of the hands, the wrists, and sometimes the legs[212]. Flat warts are caused by HPV types 3,10 and 28[213].

Deep palmoplantar warts are similar to common warts except the lesion lies deep to the plane of the skin surface. These warts are mainly found on the planter face of the toes, the roots of metatarsi, and the heel[212]. Deep palmoplantar warts are caused by HPV types 1 (most common) followed by types 2, 3, 4, 27, and 57[213].

Cystic warts tend to appear on weight-bearing surfaces like the sole and have a smooth appearance. Cystic warts are caused by HPV type 60[213].

Focal epithelial hyperplasia includes warts that occur in the oral cavity. These small lesions appear as whitish papules, measuring 1-5 mm in size and arranged in groups. Focal epithelial hyperplasia is caused by HPV types 13 and 32[213].

Butcher's warts are seen in individuals who handle raw meat products. These warts tend to have a cauliflower-like appearance and tend to be extended. Butcher's warts are caused by HPV type 7[213].

Anogenital warts (AGW), also known as *Condyloma Acuminatum* (CA), are a benign cellular proliferation of the anogenital skin and mucosa in response to an HPV invasion, and there are hundreds of thousands of new cases every year. The number of new cases of genital warts registered every year in the USA is 350,000[193,214] and around 750,000 in Europe[215]

Anogenital warts are mainly transmitted during sexual intercourse and seem to be more frequent in males. They are mainly found on the foreskin of the glans, and the ridge of the foreskin in men;

on the clitoris and the large and small lips of the vulva in women. They may spread out to the mucous membranes in the anal canal, vagina, and cervix[212].

Regarding anogenital warts, human papillomavirus types 6 and 11 are considered responsible for 90 per cent of the cases[216].

As described here, some studies found a specific relation between different kinds of clinical and histological manifestations and particular types of HPV causing warts.

However, since HPVs are found ubiquitously on the human body, the identification of the ones causing the specific diseases is challenging, and these correlations need to be confirmed[217].

Finally, both cutaneous and mucosal warts are common manifestations in immunocompromised patients, such as HIV positive individuals and renal transplant recipients[218–223].

4.2 Epidermodysplasia verruciformis

Epidermodysplasia verruciformis (EV) is an autosomal recessive dermatologic condition in which patients show a decreased immunologic ability to defend against certain types of HPVs. These last are facilitated in establish persistent infections, often leading to the occurrence of verrucous cutaneous lesions such as multiple persistent verrucae, pityriasis Versicolor-like lesions, and other verrucous or "warty" cutaneous lesions as well as in the development of Bowen disease and squamous cell carcinoma[196,224,225]. Patients in this condition are more prone to develop cutaneous lesions. 50% of patients with EV develop skin cancer by age 50[225].

EV is a sporadic disease, but its relation with HPV infection and the high incidence of the malignant outcome, kindle the interest of the scientific community. The study of this disease can help the comprehension of viral infections and their role in carcinogenic pathways[224].

There are two forms of EV, a primary type, inherited in an autosomal recessive pattern and a secondary one, clinically indistinguishable, observed mainly in HIV-infected, solid organ transplant recipients, both immunocompromised, or immunosuppressed individuals[224,226–228].

The main HPV types associated with EV (EV-HPV) and cancer occurrence are HPV5, HPV8, and HPV14[225]. Many other HPV types are related to EV disease but not with cancer occurrence. In general, the role of EV in the onset of cancer is not clear and still debated[229].

4.3 Head and neck cancer

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide, with an estimated annual burden of 355,000 deaths and 633,000 incident cases[230].

HNSCC can arise from mucosal epithelial cells from the oral cavity, oropharynx, hypopharynx, larynx, sinonasal tract, and nasopharynx.

The majority of HNSCC is keratinizing or non-keratinizing squamous cell carcinoma, ranging from well-differentiated to undifferentiated types[231,232]. The higher incidence of these cancers has been registered in white men in the age range 40-55. HNSCCs are etiologically heterogeneous, being caused by tobacco use, alcohol consumption, poor oral hygiene, exposure to certain chemicals, and genetic features[233–235], as well as viral infections[232,236]. HR HPV infections, particularly with HPV16[237], have been associated with a subset of HNSCCs, and in particular, with oropharyngeal cancer (OPC). HPV16 infection accounts for the 90% of the HPV-related OPCs[31,238].

While the overall incidence of HNSCC is decreasing, consistently with declines in tobacco use, an increased incidence of oropharyngeal cancers related to HPV infection in younger people with no history of alcohol abuse or tobacco consumption has been registered[239–241]. More than 19,000 new oropharyngeal cancer cases associated with HPV infection are registered in the USA every

year[242]. The estimated number of annual new cases of oropharyngeal cancer in Europe is close to 100,000. The estimated proportion of oropharyngeal cancers attributable to HPV for specific geographical regions is 56% in North America, 52% in Japan, 45% in Australia, 39% in Northern and Western Europe, 38% in Eastern Europe, 17% in Southern Europe, and 13% for rest of the world [241,243].

A correlation between the number of oral sex partners and the risk of oropharyngeal cancer has been reported[244].

Oropharyngeal cancers associated with HPV infections seem to have a better prognosis than HPV-negative tumors, with better response to chemotherapy and radiation therapy[232].

HPV-associated oropharyngeal cancers arise most commonly from the lingual and palatine tonsils[245]. It seems that HPV preferentially targets the highly specialized reticulated epithelium of the tonsillar crypts. Still, the intrinsic properties of this epithelium that make it vulnerable to HPV infection are not clear[246].

In HPV-positive oropharyngeal cancer, key events are the p53 degradation[247], pRb pathway inactivation[248], and p16 upregulation[249]. In contrast, oropharyngeal cancers related to tobacco consumption are characterized by p53 mutation, downregulation of p16, and pRB upregulation[250]. In this context, immunohistochemistry for p16 can be used as a surrogate marker[232] and HPV test to predict the likely or expected development of the disease (Figure 11). Oropharyngeal cancers often occur without pre-cancerous manifestations, causing a late identification of the malignancy and belated treatments. Survival rates range from 40 to 50%, and this is probably due to late diagnosis[218].

In a study conducted in India in 2017, the contribution of HPV infection in head and neck cancers was evaluated.

This work indicates that the proportion and types of mucosal HR-HPV associated with HNC in India differ from those in other more developed parts of the world. Differences in smoking and sexual behaviour between India and North America and northern Europe may underlie these variances. Moreover, this work showed that $p16^{INK4a}$ staining appeared not to be an excellent surrogate marker of HPV transformation in the Indian HNC cases[239].

In patients like HIV-positive individuals, the immunocompromised status can favour the HPV infection of the oral cavity[251]. In particular, HIV positive patients are three times more prone to HPV infection of the oral cavity compared to healthy individuals[252].

Among the oral HPV infections in HIV positive patients, HPV16 is the most frequent papillomavirus usually detected[232,253].

The introduction of antiretroviral therapies (ART) in the late nineties resulted in a dramatic improvement of clinical outcomes and life expectancies for people living with HIV.

The improved immunological status achieved thanks to the ART therapies, was supposed to produce also an inhibitory effect on the HPV infections of the oral cavity, being the immune system more responsive, but results are not clear[253,254].

Transplant recipients are also susceptible to contract oral HPV infection and develop HPV-related oral cancers[253]. In these patients, oropharyngeal carcinoma is the third most common HPV-related cancer, after vulvar and anal carcinoma[255].

Hematopoietic Stem Cell Transplantation (HSCT) recipients are also particularly susceptible to oropharyngeal cancer development, and the HPV infection seems to play a role in this process[253].

Many studies suggest that HPV copy number is low in oropharyngeal cancers. HPV DNA was found, using in situ hybridization, in scattered foci of epithelia, indicating a non-clonal origin of the tumor. The hypothesis is that HPV replication occurs only in some infected cells.

Differently from other oropharyngeal cancers, the copy numbers of HPV in tonsillar carcinomas display a wide variation[256]. A study conducted by Koskinen and colleagues detected a higher median HPV copy number in tonsillar specimens compared to other non-tonsillar oropharyngeal cancers[257].

Analyses on tonsillar carcinomas showed an HPV copy number between 10 and 100 copies per one beta-actin, and patients with tumors containing >190 copies per beta-actin showed a significantly better clinical outcome[256].

Additionally, patients with episomal viral DNA, more frequently have large (T3–T4) oropharyngeal tumors than patients with integrated or mixed forms of viral DNA[257].

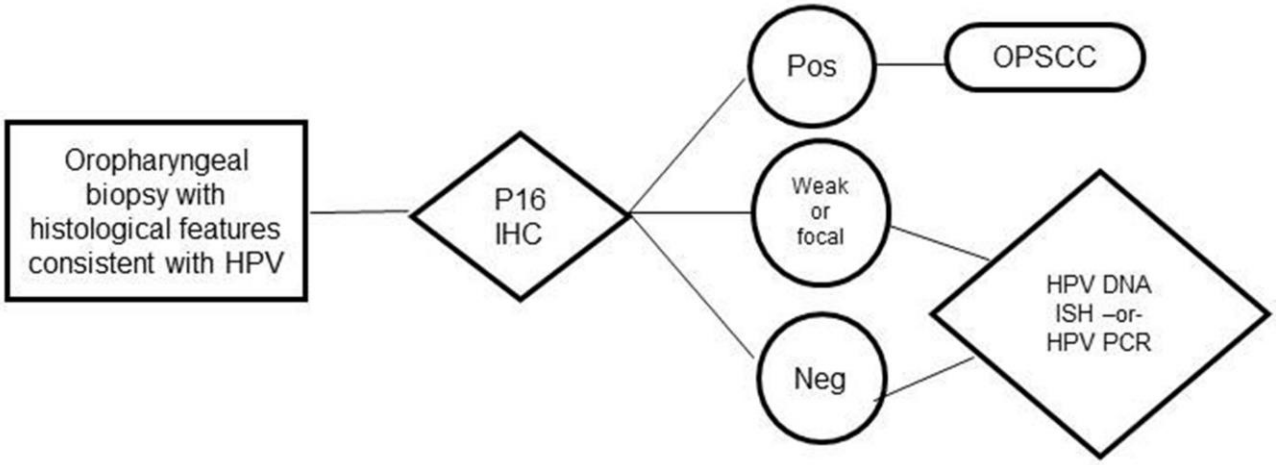


Figure 11: Correct procedure for the diagnosis and follow-up of the oropharyngeal cancer[218]

4.4 Anal cancer

Anal cancer is a very rare malignancy with low incidence rates (2/100,000 for both men and women)[258].

In the last few years, its incidence increased in men who have sex with men (MSM) and in HIV positive individuals[259–263]. High-risk HPV and HIV coinfection are considered a risk factor for anal cancer[264]. More than 90% of anal cancers are associated with persistent HPV infection, and the majority of these cases are attributed to HPV16 and HPV18[264,265].

In men and women with HIV, the prevalence of anal HPV detected surpasses 90%, though not all will have abnormal pathology[218].

In HIV positive patients, the incidence of anal cancer is 30-fold higher compared to healthy individuals. In men who have sex with men (MSM), this value increases to 80-fold[260,264]. The increased lifespan of HIV positive people, due to the development of effective therapies, is associated with an increase of anal cancer cases in this cohort of patients[260].

Also, immunocompromised solid organ transplant patients are more susceptible to anal cancers with an increased risk of 10-fold.

There are currently no formal recommendations for routine anal cancer screening, but general algorithms are present[266] (Figure 12).

There are sub-groups of patients like men having sex with men that are more encouraged in following screening programs. In those particular cases, HPV testing could be a useful prevention method; however, as for the cervical cytology, also anal cytology results of difficult interpretation[267].

Anal Papanicolau test and high-resolution anoscopy (HRA) with biopsy represent the gold standard for the identification of anal cancers[218,268,269].

Precancerous anal lesions are termed anal intraepithelial neoplasia (AIN) of grade 1, 2, or 3 (mild, moderate, or severe). The p16 immunohistochemistry (IHC), a well-validated surrogate marker of HPV transformation in the cervix[270–273], is recommended to determine the appropriate lesion classification[267] (Figure 12). In 2018, Donà and colleagues evaluated the HPV status and the

presence of anal lesions in HIV–infected and HIV-uninfected men who have sex with men (MSM). Both HIV-uninfected and HIV-infected participants positive at least for one HR (i.e., HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, or 68) and at least one LR (i.e., HPV6, 11, 40, 42, 54, 61, 72, 81, or CP6108) types showed significantly increased odds of having LSIL+ in comparison with those with only HR types. Thus coinfection with HR and LR types could affect the anal lesions formation[274].

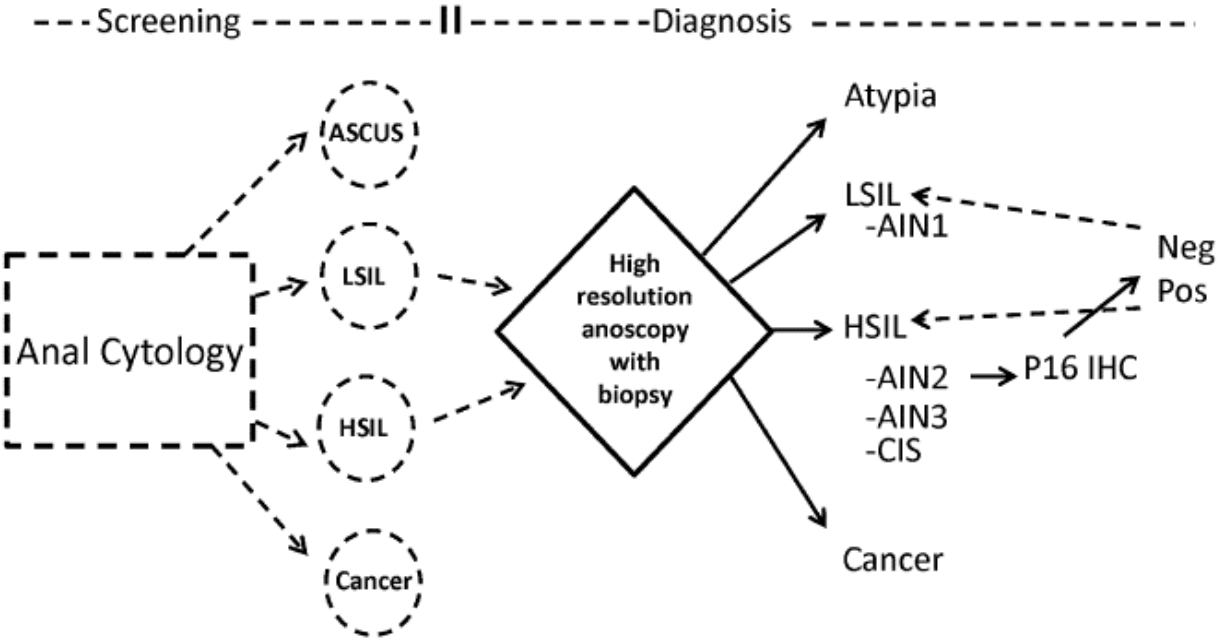


Figure 12: Correct procedure for the screening, diagnosis, and follow-up of anal cancer[218]

4.5 Penile cancer

Penile cancer is a sporadic disease with an incidence of 0.5% in the male population[275,276]. Approximately 50% of penile cancers are HPV positive, and HPV16 is the most common type related to the disease[277,278]. HIV patients have a 2 to 3-fold increased risk of developing penile cancer. Around 4% of HIV positive patients have a precursor lesion, but only 5-30% of these cases progress to cancer. Other risk factors include tobacco use, poor hygiene, phimosis, and

lack of circumcision[279]. Early diagnosis is based on the recognition of anatomical anomalies, like ulcers or other lesions. Flat penile lesions, in particular, seem to be associated with HPV infection[280].

Shear wave elastography, fluorodeoxyglucose-PET with MRI and ultra-small paramagnetic iron oxide enhanced MRI, are used for early identification of the malignancy. Moreover, the recent development of video endoscopic inguinal lymphadenectomy and robotic-assisted video endoscopic inguinal lymphadenectomy, have improved the diagnostic efficacy[276].

Biopsy histology represents the gold standard for classifying the malignancy and identifying HPV negative and positive penile cancers[281].

Most penile cancers are HPV negative and are usually classified as keratinizing squamous cell carcinomas. The ones associated with HPV infection are warty, basaloid, or mixed histologic types that show intense p16 IHC staining and the presence of lymphovascular invasion[280,282].

HPV status, morphology, and p16 IHC are used in the differential diagnosis of primary high-grade urothelial carcinoma versus HPV-related basaloid or warty/basaloid squamous cell cancer of the distal penile urethra[283]. The absence of p16 was associated with recurrence, especially in patients with lymph node involvement[284].

4.6 Cervical cancer

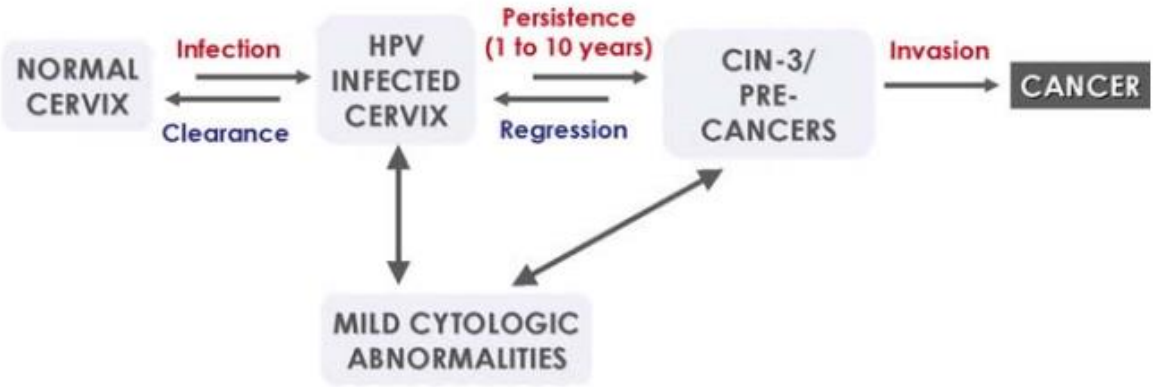


Figure 13: Major steps in cervical carcinogenesis[285].

Cervical cancer is the fourth most common cancer worldwide, with 500,000 new cases every year[201].

The pathogenesis of cervical cancer has been extensively studied in the last decades. The most critical risk factor for developing cervical cancer is infection with HR HPVs[286]. It has been established that HPV persistent infection is associated with cervical cancers[287].

Once an HR HPV type infects the cervix, this infection can be cleared by the immune system or persist for years. A longer persistence of the virus (from 1 to 10 years) correlates with a higher probability of developing high-grade lesions or even a *carcinoma in situ* (Figure 13)[285].

An exception to this rule is represented by HPV61 that can establish a persistent infection that does not flow into cancer development[288].

In most cases, the infection is cleared in a few months, and the epithelium return in his normal status. In this process, the innate immune system plays a central role.

In a small number of cases, the infection persists and leads to microscopic abnormalities. At this step, the host immune system can still eradicate the infection leading to the regression of cervical abnormalities.

In rare cases, the HPV infection persists despite the immune system, leading to the progression to precancer and invasion[285].

HPV types 16 and 18 together are responsible for the 50% of precancerous lesions of the cervix and 70% of cervical cancers[289,290].

HPV16 or 18 infections are cleared more slowly than infections caused by *othe*r high-risk types, favouring the progression to cervical lesions and cancer[291].

Risk factors for cervical cancer development are tobacco consumption, early sexual debut, a high number of sexual partners, a history of other sexually transmitted infections like HIV, and a history of cervical lesions or carcinoma[292]. A higher risk of developing HPV infections and cervical cancer were registered in women with HIV and solid transplant (SOT) recipients[196,293]. However, the incidence in HIV-positive and SOT patients undergoing modern therapies is not clear[293–295].

Cervical carcinoma is the third most common secondary malignancy in HSCT (Hematopoietic Stem Cell Transplantation) recipients[296,297]. Women who receive long-term treatment for chronic Graft-versus-Host disease (cGVHD) and those with an unrelated HLA-matched donor are at the highest risk for developing lesions after transplantation[297,298].

The HIV positivity seems to correlate with a 22-fold higher incidence of HPV infection even when the CD4 count is normal[71,299]. Although the majority of women counteract HPV infection at least once in their lifetime, most of these infections regress in a couple of years.

Since 1941, Papanicolaou-stained cervical cytology has been the standard screening test for the detection of premalignant lesions and cancer[218]. More recent technology for the sample preparation is the liquid-based cytology that requires placing cervical specimens into a vial of liquid preservative. This preparation preserves the sample and gives more reproducible results. Then, stained slides are examined manually using a standard light microscope[300].

The Bethesda System has been the standard terminology used for reporting cervical cytology since 1988 and is continuously updated.

Atypical squamous cells (ASC) are common abnormal findings that are divided into ASC-US and ASC-H. Low grade squamous intraepithelial lesion (LSIL) is a mild dysplasia caused by HPV infection, characterized by squamous cells with nuclear atypia, perinuclear halo, and dense cytoplasm. High grade squamous intraepithelial lesion (HSIL) is a more severe abnormality that includes moderate to severe dysplasia and carcinoma in situ. An older classification method describes three levels of pre-cancerous lesions, CIN 1, CIN2, and CIN3, with a progressively

higher amount of cellular aberrations and tissue abnormalities. The Bethesda system also allows for the description of glandular cell abnormalities, which are caused by HPV but are far less common and are categorized as atypical glandular cells (AGC). “Adenocarcinoma in situ” and “AGC favour neoplastic” are also included as subcategories of AGC.

Human-papillomavirus (HPV) DNA testing is now used as an alternative to primary cervical cancer screening using the cytological examination.

In the past, cervical cytology has been very useful in cervical cancer screening programs but carries a false-negative rate of about 5% to 20%[301]. Thus, HPV DNA testing has been introduced in cervical cancer screening programs to compensate for the limitations of the cytology tests[218].

In a study conducted to evaluate the effectiveness of HPV DNA testing for cervical cancer screening, the relative detection, relative specificity, and relative positive predictive value (PPV) of HPV DNA testing *versus* cytology were assessed. Overall evaluation of relative detection showed a significantly higher detection of CIN2+ and CIN3+ for HPV DNA testing *versus* cytology. The cytology had higher relative specificity in detecting both CIN2+ and CIN3+ lesions, considering all age groups. In contrast, HPV DNA and cytology tests had similar specificity in detecting both CIN2+ and CIN3+ lesions, considering just women of age 30 or older. Therefore, primary screening of cervical cancer by HPV DNA testing appears to offers the maximum detection of CIN2+ and adequate specificity, if performed in the age group ≥30 years[302]. On the other hand, HPV DNA screening in women aged <30 years may lead to overdiagnosis of regressive CIN lesions, and women in this age group should not be screened with HPV DNA testing[303,304].

The high sensitivity of HPV DNA testing could only marginally improve by systematically adding the cytology test with an irrelevant increase of precancerous lesions detected, as already shown by the results of several other studies[305–308].

The high sensitivity of the HPV DNA testing has to be taken into account, especially to evaluate the procedures to be implemented in case of test positivity, to avoid overdiagnosis[302].

Several HPV DNA assays have been developed. Most of these tests have been designed for the detection of a few high-risk HPV types, including HPV16 and 18.

More recent protocols allow the detection of the HPV E6/E7 mRNA, which provides an essential advantage because E6/E7 mRNA expression correlates with increased cervical lesion severity and thus is a better indicator of disease progression than the only HPV DNA detection[309,310].

The establishment of productive infections by oncogenic types of HPV is a critical early event in the natural history of cervical cancer. There is a positive relationship between viral load and the likelihood of persistent HPV infection[311] and, consequently, also with the risk of developing cervical neoplasia[312].

Therefore, measurement of high-risk HPV types viral load, as a surrogate for HPV persistence, may identify women at risk of developing cervical cancer precursors[313].

Moreover, several studies suggest that HPV16 integration into the host genome is associated with neoplastic progression. Thus the evaluation of the HPV16 E6/E2 ratio has been proposed as a potential marker of cervical neoplastic progression[314–316].

However, the presence of both integrated and episomal HPV genomic forms in cervical cancers has been proven. Thus the evaluation of integration events cannot be used as an effective method to define the risk of neoplastic progression[314]

The different countries use different algorithms for screening, diagnosis, and follow-up. More and more countries are using HPV DNA testing as a preferred method for the first screening. In contrast, the cytology is mainly used as a complementary analysis for the diagnosis and follow-up[317] (Figure 14).

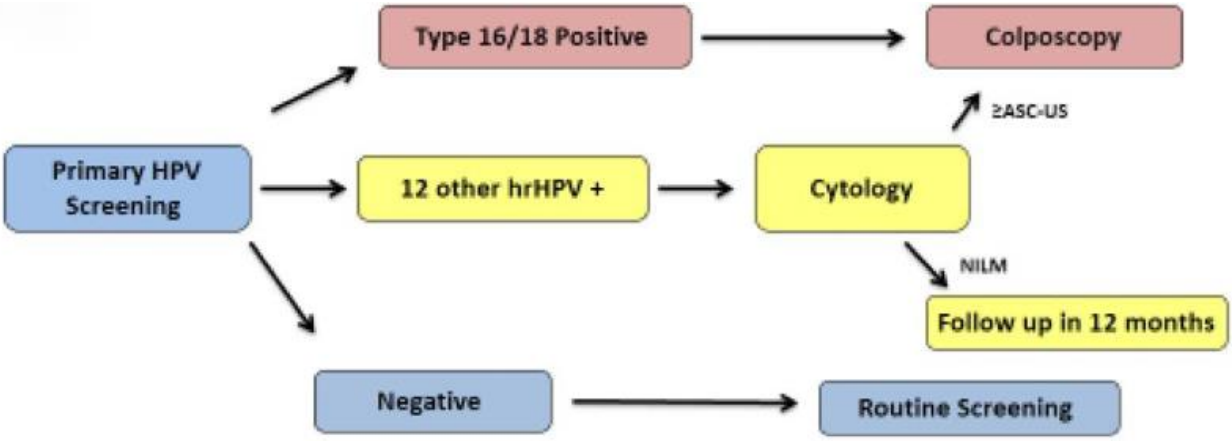


Figure 14: Algorithm adopted for the screening, diagnosis, and follow-up of cervical cancer in US[317]

4.7 Skin cancer

Skin cancer is the most common human cancer worldwide. The main risk factors for skin cancer development are UV-radiation-exposure and older age[318,319]. The incidence rate of both non-melanoma and melanoma skin cancers steadily increased over the last decades[320].

The annual incidence of NMSC reaches 2 to 3 million cases and 132,000 cases for melanoma skin cancer worldwide. A potential decrease of around 10% of the ozone in the atmosphere is estimated to cause an additional 300,000 non-melanoma and 4,500 melanoma skin cancers cases per year[321].

Also, recent studies support the involvement of beta HPVs together with UV radiation (UVR), in the development of cutaneous squamous cell carcinoma (cSCC)[55,202–207].

Two members of the genus beta, HPV5, and HPV8, were firstly detected in skin lesions of patients with epidermodysplasia verruciformis (EV) and were classified as possibly carcinogenic to humans (IARC Group 2B)[177]. In 30 to 60% of EV cases, these skin lesions can evolve into squamous cell carcinoma at anatomical sites exposed to sunlight[183,322,323].

Cutaneous squamous cell carcinoma (cSCC) is the most common malignancy developed by solid organ transplant recipients (OTRs) with a 65-250-fold increased risk to develop cSCC compared to the general population[324–326]. Life-long immunosuppressive therapy, adopted with OTRs patients, is the most crucial risk factor for developing cSCC in these patients. Other important risk factors include sun exposure, male gender, older age, smoking, and fair skin with susceptibility to sunburn[327]. Moreover, the role of human papillomaviruses (HPVs) in the development of cSSC in OTRs has been frequently suggested[205,328–330].

In HIV patients, a 2-fold increased risk of cSCC compared with HIV-uninfected people has been registered[331,332]. The correlation between the immunodeficiency state and an increased risk of developing cSCC suggested a possible role of infectious agents[333], such as cutaneous HPV[202].

The involvement of beta HPVs in cSCC is hypothesized but challenging to prove, as beta HPV types are found ubiquitously on the skin of healthy individuals. Also, it has been shown that beta-HPV viral loads decrease with the progression to the cancer[334].

However, several epidemiological studies have reported a correlation between the presence of beta HPV DNA in eyebrow hairs, serology, and history of cSCC[202,335–341]. At the same time, no association was found with skin basal cell carcinoma (BCC)[342].

A possible scenario has been recently proposed where beta HPVs may play a role at the beginning of the carcinogenesis facilitating the accumulation of UVR associated mutations and aberrations induced by UV radiation[343]. At the beginning of the process, mutations can occur in key genes such as p53. This event can facilitate the loss of proliferative control and the progression of the transformed phenotype (Figure 15).

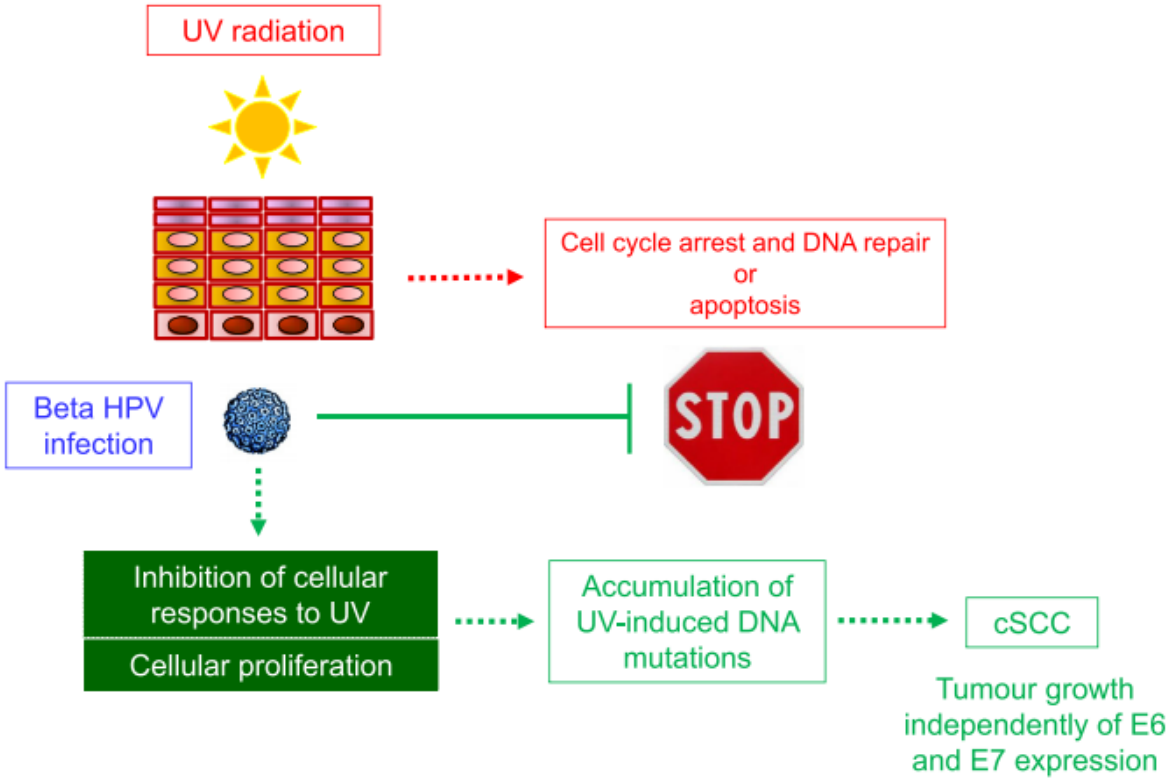


Figure 15: Proposed model describing the potential role of beta HPVs in cooperation with UVR in cSCC development[203].

In vitro experiments showed the transforming abilities of E6 and E7 proteins from beta HPVs in different cell lines and primary keratinocytes[183]. HPV38 E6 and E7 can suppress the activity of p53 and pRb, thus promoting cells immortalization[344,345].

In another study, the properties of E6 and E7 oncoproteins from six uncharacterized beta HPVs (i.e., 14, 22, 23, 24, 36, 49) were compared. Only HPV49 E6 and E7 immortalized primary human keratinocytes and efficiently deregulated the p53 and pRb pathways. Instead, the other beta types (i.e., 14, 22, 23, 24, 36) were only capable of extending the lifespan of primary human keratinocytes without inducing cells immortalization. Moreover, HPV49 E6, similarly to E6 from the oncogenic HPV16, promoted p53 degradation[189].

Other studies showed that beta HPV5, 8, and 38 were able to alter the cell cycle, DNA repair, apoptosis, and activation of immune-related pathways[183,340].

E6 from several beta types, including HPV8, 25, 98, 17A, 38, 76, and 92, inhibits the Notch pathway, through the interaction with Mastermind-like 1 (MAML1), delaying the keratinocyte differentiation[346–349].

Also, studies on mouse models demonstrated the ability of beta HPV8 and 38 E6 and E7 in promoting cSCC upon initialization with UV irradiation[343,350,351].

A mouse model expressing HPV38 E6 and E7 specifically in skin keratinocytes was developed and showed higher susceptibility to cancer formation upon UV irradiation. Several genes were mutated, in particular p53 and Notch. The silencing of the expression of the viral genes in established skin lesions didn't affect further tumour growth. Once p53, and likely other cellular genes, were irreversibly inactivated by DNA mutations induced by UV radiation, the progression, and maintenance of the carcinogenic skin process seamed to become independent of the expression of viral genes. In contrast, the loss of the viral genes at the early stages of the irradiation protocol prevented the development of UV-induced skin lesions, underlining the key function of HPV38 E6 and E7 in the early stages of the UV-mediated carcinogenesis[343].

Another study, using a mouse model infected with murine PVs sharing functional similarities with human beta HPV types, confirmed the potential role of HPV types in skin cancer development. As previously shown, in human cancers, there is a substantial decrease of viral load in the late stages of the malignancy, confirming the theories about the role of beta HPVs in early stages on skin lesions[351].

Altogether these studies support the hypothesis that beta HPVs have developed strategies to maintain infected cells in a proliferative status to efficiently complete their life cycle in the skin, leading to the accumulation of UVR-induced mutations and increased risk for skin cancer.

5 Molecular tools for the discovery of new HPVs

The discovery of new HPVs has followed the evolution of the technologies. During the last few years, the number of new HPV types has grown dramatically due to the arrival of high throughput sequencing methods (Figure 16).

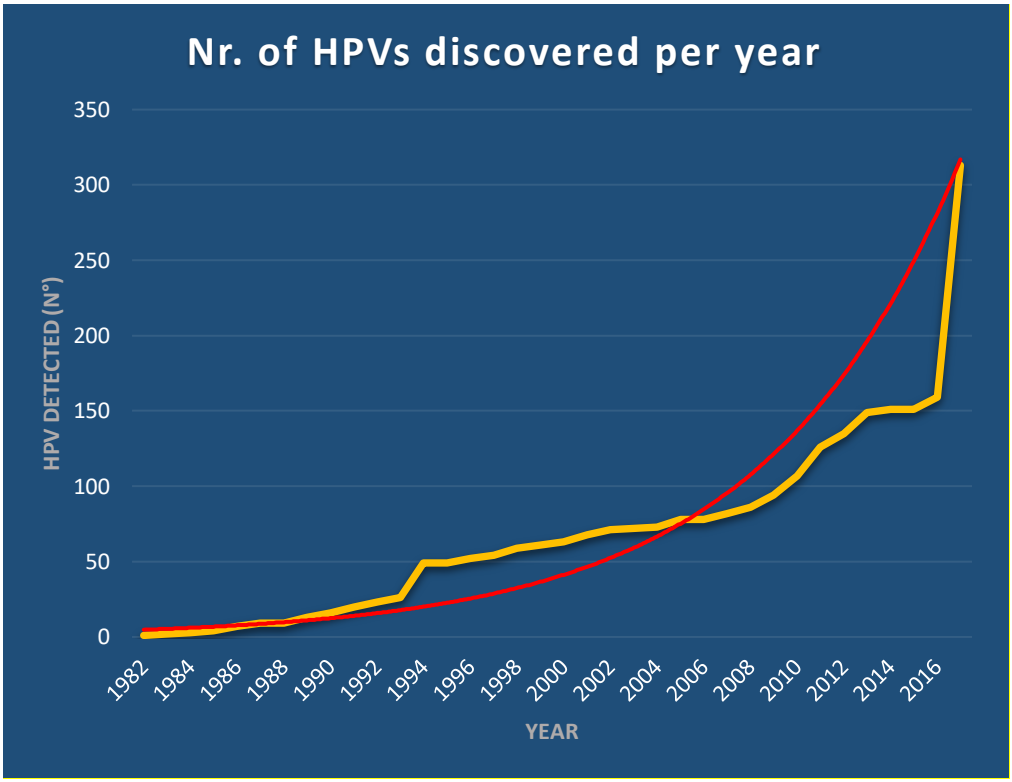


Figure 16: Number of new HPV types discovered per year according to the data from the Papillomavirus genome database (PaVe), updated in October 2017. In yellow the line representing the HPV discovery and in red the trend line that is exponential[352]

5.1 PCR primers systems for the detection of HPVs

The polymerase chain reaction (PCR) represents the first method used to detect and identify HPV sequences in different samples[353–355,355,356]. The L1 ORF is used to characterize the different HPV types, and several consensus or degenerate primers sets have been developed so far targeting this region [357].

The MY09/MY11[358,359] and the GP5/GP6[360] were developed around the beginning of the 1990s and represent the first and widely used primer sets for the detection of HPV DNA sequences. These primer sets were initially designed for the identification of mucosal HPV types in the human genital tract[361].

MY09-MY11 primers set, developed in 1989, is composed of a mix of 25 degenerate primers capable of amplifying a broad spectrum of HPV types[359].

In 1990 the GP5/GP6 primer system was developed. This system is composed of GP5 (5'-TTT GTT ACT GTG GTA GAT AC-3') and GP6 (3'-ACT AAA TGT CAA ATA AAA AG-5') couple of general primers, which span a region of 140-150 bp from the L1 open reading frame representative of 6 HPVs (i.e. HPV6, 11, 16, 18, 31 and 33)[360]. These general primers showed to amplify a broather spectrum of HPV types (i.e. HPV 6, II, 13, 16, 18, 30, 31, 32, 33, 35, 39, 40, 42, 43, 44, 45, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61 and 66) under conditions that allow mismatch acceptance[362,363].

In 1991, Snijders et al., designed two degenerate primers GP17 and GP18 and one general probe GPR22, targeting the L1 ORF of HPVs. The effectiveness of the primers and the probe was tested on a panel of 24 cloned HPV DNAs, isolated from cutaneous and mucosal lesions, including HPV-2a and -57, which are known to be associated with lesions at both anatomical sites.

This primers set was able to amplify most of the HPV types tested, namely pHPV-1a, -2a, -5, -6b, -7, -8, -11, -13, -18, -25, -31, -32, -33, -38, -39, -41, -43, -45, -46, -51, -56 and -57.

In 1994 a new PCR assay for the detection of HPV sequences was developed. A study of the sequences of 45 HPVs and 9 animal PVs, known at that time, allowed the identification of conserved sequences located in the L1 ORF[364]. These regions happened to coincide with the primer regions previously described by Snijders and colleagues in 1991[365]. The sequences of the original GP17 and GP18 primers, as well as the sequence of the oligonucleotide GPR22 used as

the degenerate probe, were modified to cover a broader spectrum of known PVs and also to attempt to detect unknown, distantly related PV types.

The use of these degenerate primers generated some unspecific amplification, raising concerns about the specificity of the assay. DNA sequencing of the obtained PCR product was used to verify the specific amplification of PVs sequences.

In 1995, a nested-PCR protocol was developed for the amplification of E1 and L1 HPV sequences based on the alignment of 19 different mucosal HPV types[366]. The use of a nested-PCR approach showed to be effective in reducing the unspecific amplification of non-viral sequences present in the samples. This new nested-PCR protocol showed a better sensitivity, compared to the MY09/MY11 protocol of 1989, in detecting different HPV types in the genital tract (i.e. HPV6, 11, 16, 18, 30, 31, 33, 34, 35, 39, 40, 42, 45, 51, 52, 53, 56, 57, and 58).

In 1995, Berkhout and colleagues developed a new nested PCR protocol, based on the use of degenerate primers, enabling the detection of known EV-associated HPV types with high sensitivity[367].

These degenerate primers revealed many multiple infections never characterized before. An improved version of these primers was subsequently developed by Boxman and colleagues, allowing the detection of a wide range of cutaneous PV types[368].

In 1995 the GP5+/GP6+ primer system, an improved version of the original GP5/GP6 primer system, was generated. Alignment of the L1 region from 24 mucosotropic HPVs and the sequences from GP5 and GP6 primers allowed the identification of the consensus sequences Thr-Arg-Ser-Thr-Asn (TRSTN) immediately downstream of the GP5 (forward primer) region and Arg-His-X-Glu-Glu (RHXEE) upstream of the GP6 (backward primer) region[369]. These amino acid conservations reflect codon conservations at the nucleotide level[370]. Part of these conserved sequences was used to elongate GP5 and GP6 at their 3' ends to generate the primers GP5+

and GP6+, respectively. Compared with the GP5/6 PCR, GP5+/6+ specific PCR on 22 cloned mucosotropic HPVs (i.e. HPV6, 11, 13, 16, 18, 30, 31, 32, 33, 35, 39, 40, 43, 45, 51, 52, 54, 55, 56, 58, 59 and 66) revealed an improved HPV detection[370].

In 1996, Shamanin and colleagues analyzed tumors, obtained from both immunosuppressed and non-immunosuppressed patients, for human papillomavirus (HPV) DNA using degenerated primers, combined into 16 different PCR protocols, all amplifying a portion of the L1 ORF. The putative HPV sequences were cloned and sequenced to verify the specificity of the amplification and to characterize the HPV types present in the samples[371].

In a study of 1997, the PCR procedures described by Shamanin and collegues[371] and Berkhout and collegues[367] were used to amplify different DNA samples, including some of the ones previously analysed in the studies by Barr and collegues[372] and Stark and collegues[373]. This study aimed to clarify the diversity of published results for prevalence and spectrum of HPV types detected in benign and malignant lesions of the skin. In this study, 18 novel cutaneous HPVs associated with E*pidermodysplasia Verruciformis* (EV) were discovered, in addition to the detection of known cutaneous and EV-associated HPV types[374].

In 2000, the PGMY09-PGMY11 (PGMY09/11) primer system was developed as an improvement of the original MY09/11 primers system. This new primer system was designed to increase the sensitivity of amplification across the broad spectrum of known HPV types by using the same primer binding regions in the L1 open reading frame. Sequence heterogeneity was accommodated by designing multiple primer sequences that were combined into an upstream pool of 5 oligonucleotides (PGMY11) and a downstream pool of 13 oligonucleotides (PGMY09), thereby avoiding the use of degenerate bases that yield irreproducible primer syntheses. The PGMY09/11 primers showed a higher sensitivity in the amplification of several HPV types (i.e., HPV26, 35, 42, 52, 54, 55, 59, 66, and 73) compared to the MY09/11 primers system when tested on a set of 262 cervicovaginal lavage specimens[375].

In 2005, Baines and colleagues developed consensus-degenerate hybrid oligonucleotide primers (CODEHOP), to detect novel PVs[376]. The development of these primers was based on the concept that some HPVs are related more closely to animal PVs than to human genital or mucosal PVs.

They hypothesized that some HPV-associated tumors could be erroneously classified as HPV-negative because of an inability to detect HPV types that are phylogenetically distinct from the genital/mucosal known types. This study aimed to design broad-spectrum PCR primers for the detection of papillomaviruses from each of the six groups, discriminated generating a phylogenetic tree based on the L1 ORF of 106 PV sequences using the HPV Sequence Database (Los Alamos National Laboratory Bioscience Division, 2002), (Figure 17)[377].

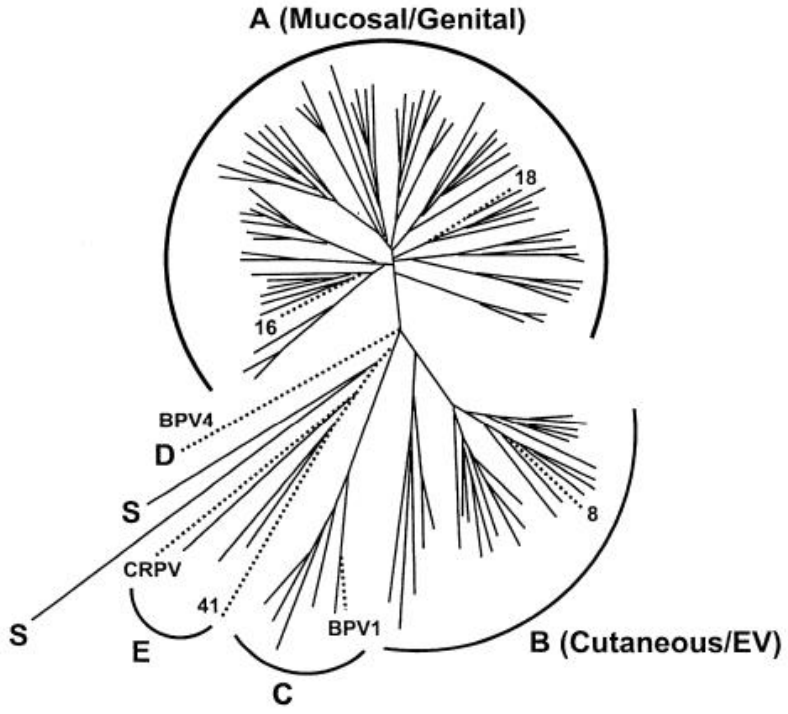


Figure 17: Neighbor-joining phylogenetic tree representative of 106 papillomaviruses based on the L1 ORF (from the Human Papillomavirus Sequence Database)[377]

According to the conserved sequences identified among these groups of PVs (i.e., A, B, C, D, E, and S), new partially degenerate primers were generated.

When compared to previous MY09/11 primers, CODEHOP primers showed to be more effective in detecting PVs from 5 of the 6 groups, when tested on plasmids and clinical samples from esophageal and tonsillar cancer.

Using these PCR primers, in 2013, more than 200 sequences representative of putative new HPV types were identified, but only a few of them have been full characterized[378]

In 2008, the BSGP5+/6+ primer system was developed, based on the original GP5+/GP6+ primers, to allow the amplification of a broader spectrum of HPVs[379].

The L1 regions of 48 completely sequenced HPV genotypes (HR HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, 82; pHR HPV types 26, 53, and 66; and LR HPV types 6, 7, 11, 13, 30, 32, 34, 40, 42, 43, 44, 54, 55, 61, 62c, 67, 69, 70, 72, 74, 81, 83, 84, 85, 86c, 87c, CP6108, 90, 97, 102, and 106) were used to develop these new primers.

The addition of eight upstream and two downstream BSGP5+/6+ (BS) primers improved amplification of plasmids of 14 genital HPV types (i.e. HPV6, 11, 16, 18, 26, 31, 33, 35, 39, 42, 43, 44, 45, 51, 52, 53, 56, 58, 59, 66, 68, 70, 73, and 82) 10- to 1,000-fold versus GP5+/6+ PCR without altering sensitivity for the 10 others.

Moreover, BSGP5+/6+ was significantly more sensitive than GP5+/6+ for detection of HPV 30, 39, 42, 44, 51, 52, 53, 68, 73, and 82, detecting 212 additional HPV infections and increasing the proportion of multiple infections from 17.2 to 26.9% in cancer patients.

This new primer system showed to be suitable for epidemiological and diagnostic applications. Besides, the integration of internal Beta-globulin PCR allowed simultaneous DNA quality control

without affecting the sensitivity of HPV detection. This protocol was later combined with Luminex technology to perform large-scale epidemiological studies[380,381].

Among the several primers systems designated to identify known and putative new HPV types, FAP[382] and CUT[383] primers, developed in 1999 and 2010 respectively, so far represent some of the most used and effective primers sets.

5.1.1 FAP primers

In 1999 Forslund and colleagues generated a single pair of broad-spectrum degenerate primers for the detection of HPVs[382].

A total of 80 HPV types L1 ORF sequences were aligned (Figure 18), leading to the identification of two relatively conserved regions, used after for the design of the FAP primers (i.e., FAP59 and FAP64), generating an amplicon of about 480 bp.

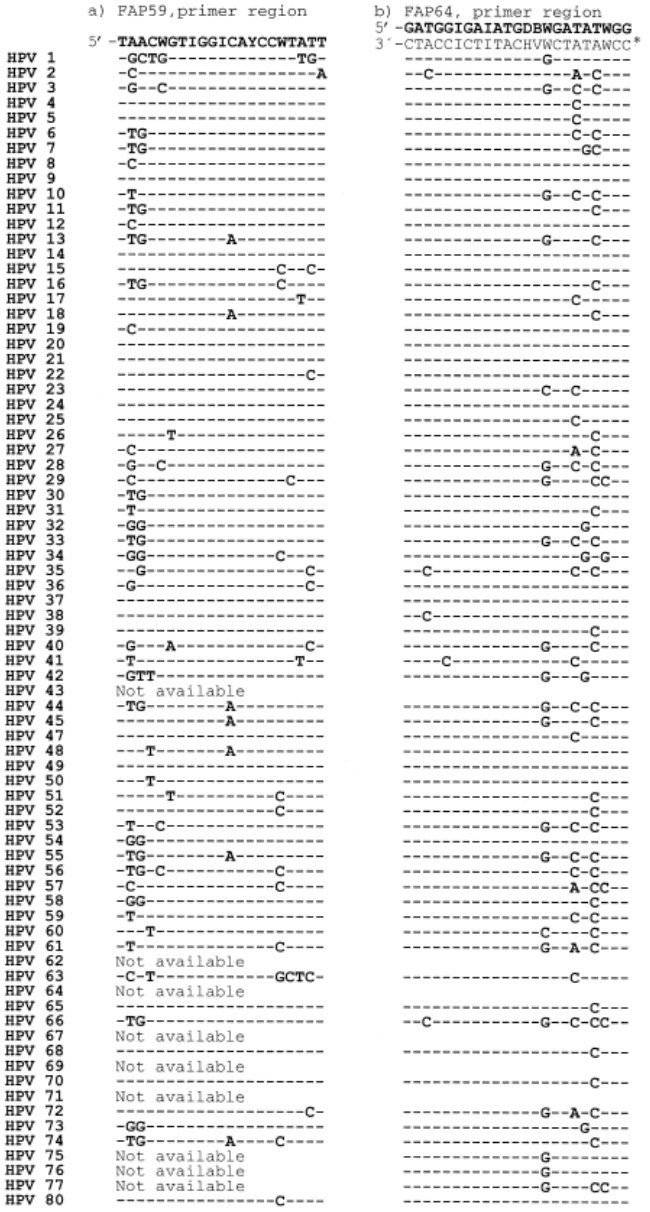


Figure 18: FAP primers alignment with 80 HPV types L1 ORF sequences: The mismatched nucleotides in each of the HPV sequences are reported. Degenerate nucleotides into the FAP primers sequences are indicated with letters: W = T, A ; I = inosine ; Y = C, T ; D = A, G, T ; B = G, C, T ; H = A, C, T ; V = A, C, G[382]

The use of these primers allowed the amplification of DNA from 87% of the HPV types tested, its sensitivity being 1-10 copies for HPV5, 20, and 30 clones (Figure 19).

These primers were tested on patients with various cutaneous tumors and healthy skin biopsies. HPV was found in 63% of tumor samples and 63% of normal skin biopsies. HPV5, HPV8, HPV12, HPVvs20-4, and six putatively novel HPV types were identified.

HPV12 and 49, as well as eight novel HPV types, were identified on skin samples of healthy individuals.

Thirty-seven percent of HPV-positive samples were found to manifest more than one HPV type. All the HPV detected manifested high similarity with HPV types associated with skin lesions and *epidermodysplasia verruciformis*. The overall HPV finding in the skin samples was 50% using the FAP primers as compared to 18% using the nested PCR protocol described by Berkhout and colleagues[367], designed for skin types. These results suggest that the FAP protocol is sensitive and generally applicable to the detection of cutaneous HPV.

A limitation of this protocol was represented by the presence of multiple infections interfering in the direct sequencing of the HPV amplicons from clinical material.

The single-round PCR system developed by Forslund and collegues[382] was capable of amplifying a broad spectrum of HPVs from both human and animal samples[182,384]. However, by using that PCR protocol, an HPV prevalence of only 26% was observed in biopsies of non-melanoma skin cancers from immunocompetent Australian patients[385].

Thus, to improve the sensitivity of this protocol, in 2003, the same author developed an improved version of the original FAP protocol, using single-tube nested 'hanging droplet' PCR, for detection of cutaneous human papillomavirus DNA of the phylogenetic group B1[385].

The 'hanging droplet' PCR showed to be effective in the amplification of HPV sequences in different skin tumors samples and also in the presence of multiple infections. Thus, this new protocol represented a valid improvement of the original FAP protocol for the detection of HPV sequences.

The advantage of this new protocol was the use of a single tube nested 'hanging droplet' PCR approach, reducing the risk of cross-contamination caused by the two PCR steps, typically required in nested PCR protocols[386].

Aligned DNA sequences from the L1 gene of characterized cutaneous HPV types of the 1996 HPV Sequence Database compendia[387] and candidate HPV 92 and HPV 93, within the region between the original FAP 59/64 primers[382], were reviewed to identify conserved regions suitable for targeting in a nested PCR.

Two regions with a relatively high degree of nucleotide sequence homology were identified and a couple of new primers, FAP 6085F (5'-CCWGATCCHAATMRRTTTGC) and FAP 6319R (5'-ACATTTGIAITTGTTTDGGRTCAA), were generated. The sensitivity of the nested PCR was increased 10-fold compared to that of the original FAP protocol when compared by testing a dilution series of a clinical sample. Moreover, the nested HPV-PCR using these primers showed to be capable of amplifying a broad spectrum of HPV types from different human samples[385].

For some samples, the FAP primers showed to be more effective in detecting HPVs compared to the new nested PCR protocol. However, the new nested PCR protocol proved to be effective in the detection of different HPV types representing a valid improvement of the original FAP protocol[385].

Even today, FAP protocol represents one of the most used broad-spectrum PCR protocols for the identification of HPV sequences and allows the characterization of both novel human[182] and animal[384,388–392] PVs.

FAP primers system has been a successful approach in studying the HPV diversity in different human samples[382,385,393,394], and hundreds of subgenomic amplicons of putative HPV types (FA sequences) from the skin and mucosal samples have been identified using this protocol[182,338,382,385,395,396].

FAP primers were also used to develop new HPV detection protocols[397,398]. For example, FAP protocol was used in combination with Luminex[399] or with next-generation sequencing (NGS)[400,401] for the detection of known and novel HPV types.

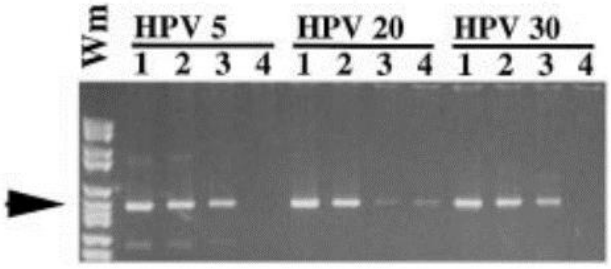


Figure 19: Analysis of PCR sensitivity with FAP59 and FAP64 primers. HPV5, 20 and 30 were amplified in a background of 100 pg placental DNA. Lanes 1–4 show 1000, 100, 10 and 1 input copies of HPV DNA in the PCR, respectively. The arrowhead indicates the position of the expected amplicon of 480 bp[382]

5.1.2 CUT primers

In 2010, Chouhy and colleagues developed new generic primers, targeting most mucosal/genital and cutaneous HPVs within all genera, and compared them to the widely used FAP primers[382,397].

Based on the L1 ORF of 88 cutaneotropic and mucosotropic HPVs, four forward primers (i.e., CUT1Fw, CUT1AFw, CUT1BFw, and CUT1CFw), and one reverse primer (i.e., CUT1BRv) were developed (Figure 20). The PCR with CUT primers generates an amplicon of around 370 bp.

The effectiveness of the new CUT and the FAP protocols were compared. Both these primers systems were used to amplify HPV sequences in 304 skin samples collected from a total of 71 patients under different pathological conditions (i.e., NMSCs, premalignant and benign skin lesions) at the Hospital Provincial del Centenario.

Overall, HPV DNA was present in 55% of the samples. HPV DNA was detected more often by FAP than by the CUT primer system (128/304 vs. 94/304, respectively), and only 55 samples were HPV-positive by both. Moreover, 137 samples were negative for HPV DNA using both primer systems.

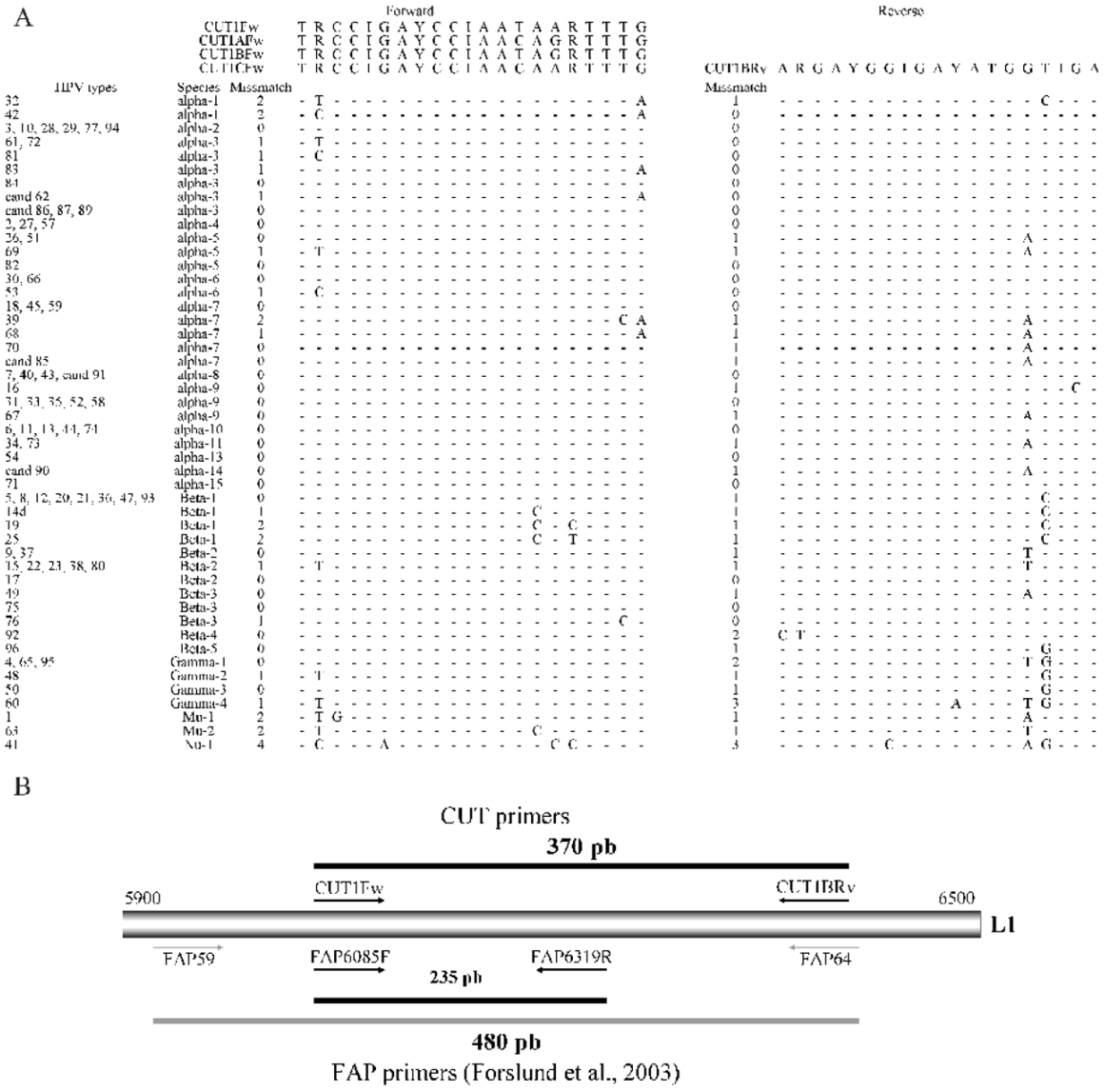


Figure 20: CUT generic primer system: A) Alignments of the CUTprimer sequences with the L1 ORF of 88 selected HPV types. B) Schematic positions of CUT, FAP59/64 and nested FAP-primers in the L1 ORF of HPV10[383]

DNA sequencing was performed to identify the HPV types detected by each primer system. Overall, 69 different HPV types/putative types were identified with both primer systems (Figure 21).

CUT primer system identified 22 different HPV types or putative types, 9 of them (41%) being novel putative types (GC05, GC06, GC08, GC09, GC10, GC11, GC12, GC13, GC14). FAP primer

system detected 37 different HPV types or putative types, 5 of which (14%) corresponding to novel putative types (GC01, GC03, GC15, GC16, GC17).

The differential capacity of CUT primers in detecting novel putative viruses compared to FAP primers showed to be significant ($p<0.01$). Finally, only ten different types were detected by both primer systems, 3 of them representing novel putative types (GC02, GC04, GC07).

This work also allowed the identification of a novel HPV type that was subsequently fully characterized and designated HPV115.

Overall, CUT primers system showed to be effective in detecting HPVs belonging to different genera and species, and represents a valid tool for the detection of known and novel HPV types. CUT primers system is currently used in different prevalence studies[402], also in combination with other techniques, such as NGS[403].

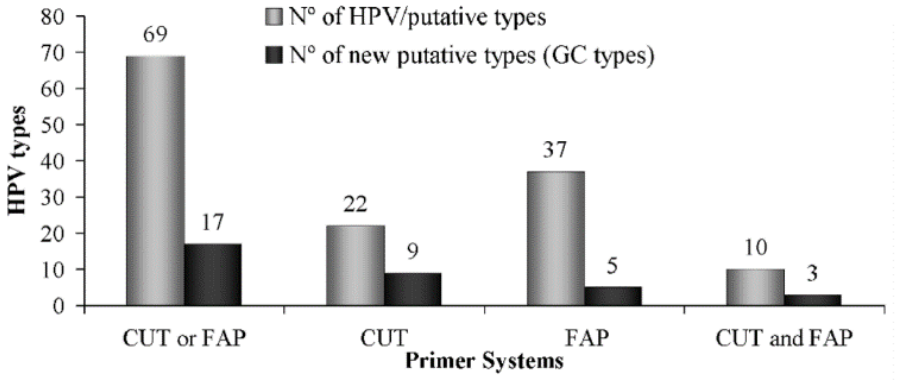


Figure 21: The differential capacity of CUT and FAP primer systems for the detection of known and novel HPV types[383]

5.2 Rolling circle amplification (RCA)

In the past, the discovery of new viruses required the use of many laborious protocols involving cell culture, density gradients, cloning, and sequencing steps[404].

Other procedures involving the use of low-stringency Southern blot hybridization, allowed the identification of HPV16 from cutaneous and condylomatous warts[405].

A portion of the genome of a putative novel HPV can be used to reconstruct the whole genome of the virus, using inverse PCR. One limitation of this strategy is the identification only of viruses related to known sequences[404].

Rolling circle amplification (RCA) is a novel technology enabling the amplification of any circular single- or double-stranded DNA molecule. This technology uses the bacteriophage phi29, a high-fidelity enzyme, with a strong strand-displacing capability, high processivity (>70,000 bases per binding event), and proofreading activity [406](15), in combination with random hexamers (Figure 22)[404].

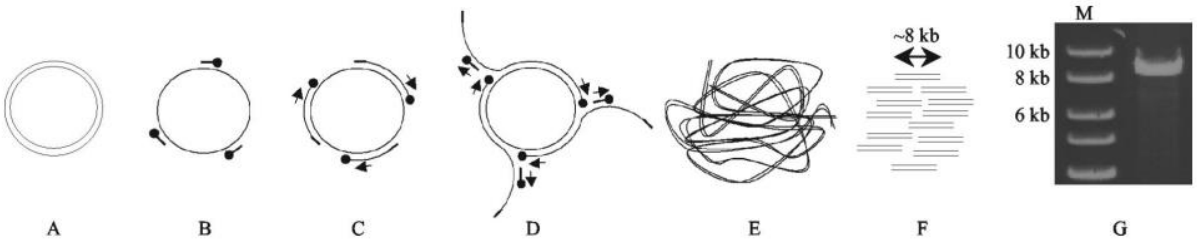


Figure 22: Schematic representation of the multiply primed RCA method for the amplification of circular PVs genomes. The first step of the process is the denaturation of the viral dsDNA into ssDNA molecules. Random hexamers can now anneal to multiple sites on this template DNA, after which the phi29 DNA polymerase binds (B) and isothermally extends these primers at the 3' end (C). Strand displacement synthesis occurs when the DNA polymerase reaches a downstream extended primer, and hexamer primers can anneal to the displaced single-stranded product strands and will again be elongated by the 29 DNA polymerase (D). The exponential amplification of the template DNA generates linear double-stranded, high-molecular-weight repeated copies of the complete PV genome (E). Digestion of this RCA product with a restriction enzyme which has only a single recognition site in the PV genome will result in multiple double-stranded, linear copies of the PV complete genomic DNA (F), which can be visualized using agarose gel electrophoresis (G)[404]

The multiply primed RCA method for amplification of circular PVs genomes is based on the use of random hexamers that bind the circular target DNA, after denaturation that generates ssDNA molecules. Thus, the phi29 DNA polymerase binds and isothermally extends these primers at the

3’ end. When the DNA polymerase reaches a downstream extended primer, strand displacement synthesis occurs, and hexamer primers can anneal to the displaced single-stranded product strands that will again be elongated by the phi29 DNA polymerase. The exponential amplification of the template DNA generates linear double-stranded, high-molecular-weight repeated copies of the complete PV genome. The digestion of this RCA product with a restriction enzyme, which has only a single recognition site in the PV genome, will result in multiple double-stranded, linear copies of the PV complete genomic DNA, which can be visualized using agarose gel electrophoresis(Figure 22 G)[404].

RCA was applied to amplify the complete circular double-stranded DNA of different PVs, without any need for prior knowledge of their sequences[404].

Moreover, used in combination with genus-specific primers, RCA can be directed towards specific PV types[407].

After the RCA amplification, restriction analysis, cloning, and Sanger sequencing are required for the identification of a novel HPV[408]. One limitation of the RCA technique is that the host genome, albeit to a lesser extent, can be amplified together with the circular viral genome, due to RCA’s ability to amplify also linear DNA molecules. Therefore, pretreatment with exonuclease or separation of the targeted DNA fragment by gel electrophoresis must be applied[409]. The use of RCA protocols, in combination with restriction enzyme analyses and qPCR, allowed the detection of HPV16 in human keratinocytes[404].

5.3 The sequencing technologies

All the information necessary to develop and maintain an organism is contained in his genome.

In 1869 the chemist Friedrich Miescher first identified what he called "nuclein" inside the nuclei of human white blood cells. The term "nuclein" was later changed to "nucleic acid" and eventually to "deoxyribonucleic acid," or "DNA.". Miescher aimed to characterize protein components of leukocytes, but he discovered the DNA instead[410].

During the following years, the contribution of many scientists like Phoebus Levene and Erwin Chargaff paved the way for the characterization of the DNA molecule. Finally, in 1953, Watson and Crick, following the crystal structure studies conducted by Franklin and Wilkins, obtained the three-dimensional, double-helical model for the structure of DNA[411,412].

The methodologies used for the study of the proteins came much before the ones used for the nucleic acids, and those techniques were not suitable to analyze this large and redundant molecules[413]. Thus it was necessary to develop new methodologies.

In 1965, R. Holley determined the complete nucleotide sequence of an alanine transfer RNA, isolated from yeast[414].

In the same period, Frederick Sanger developed a two-dimensional fractionation method, that uses radiolabeled partially digested fragments, for the determination of the nucleotide sequence with spleen phosphodiesterase [415]. This work allowed the identification of other ribosomal and transfer RNA sequences[416–420].

In 1972, using the method developed by Sanger, Fiers obtained the complete sequence of the bacteriophage MS2 coat protein gene[421], and just a few years later also the whole genome[422].

In 1977, Sanger developed the ‘chain-termination’ technique[423] that makes use of chemical analogs of the deoxyribonucleotides (dNTPs), the monomers of DNA strands. Dideoxynucleotides (ddNTPs) lack the 3′ hydroxyl group thus cannot form a bond with the 5′ phosphate of the next dNTP. Once the ddNTP is incorporated, the extension of the DNA is terminated[424]. In the reaction, both normal dNTPs and radiolabeled ddNTPs are present. The concentration of ddNTPs in the

reaction is lower than the one of normal dNTPs, and this causes the generation of DNA fragments of different lengths with the ddNTPs always in the last position of the sequence (Figure 23).

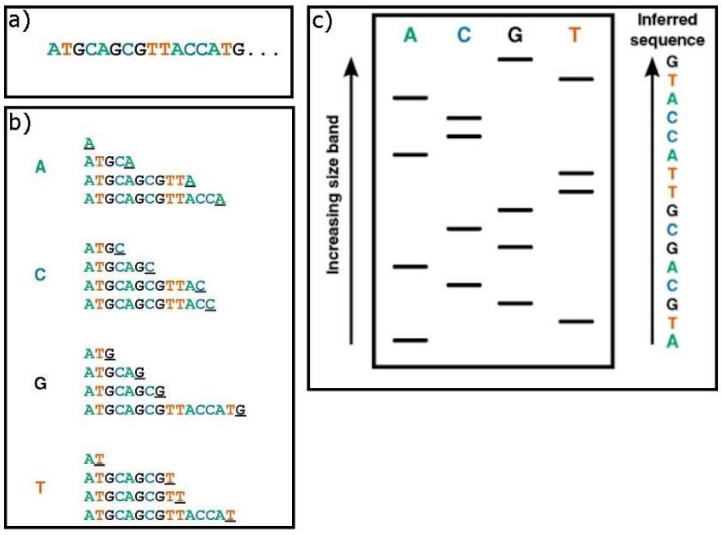


Figure 23: Sanger sequencing technology. a) Example of DNA sequence. b) Sanger "chain-termination" sequencing. The incorporation of a radiolabeled ddNTP interrupts further extension. c) The sequence is inferred by finding the lane in which the band is present for a given site. The 3′ terminating labelled ddNTP corresponds to the base at that position[425]

Four parallel reactions are performed for the four bases and loaded on an electrophoresis gel. With the use of autoradiography and by discriminating the sizes of the fragments, the nucleotide sequence of the original template is obtained (Figure 23 c).

In the following years, many improvements were made to Sanger sequencing. Particularly noteworthy is the replacement of phospho- or tritrium-radiolabelling with fluorometric based detection, allowing the reaction to occur in one vessel instead of four and the introduction of capillary-based electrophoresis, leading to the development of newer and more effective machines[425–431] up to commercial DNA sequencing machines[432].

In 1987, Applied Biosystems introduced AB370, the first automated, capillary electrophoresis (CE)-based sequencing instrument, followed in 1998 by the AB3730xl. These instruments became the primary workhorses for the NIH-led and Celera-led Human Genome Projects [433].

These first-generation DNA sequencing machines could produce reads of less than one kilobase (kb) in length.

With the advent of the "Shotgun" sequencing, overlapping DNA sequences were obtained by cloning and sequencing shorter fragments of the original long sequence, and then assembled into one long contig in silico[434,435].

After, the introduction of the polymerase chain reaction (PCR)[436,437] and recombinant DNA technologies[438,439] allowed the production of a higher amount of high-quality input DNA required for sequencing, and this further aided the genomic revolution[425].

In the next years, a great effort has been made to improve sequencing technology. Applied Biosystems produced ABI PRISM, a new sequencer that allowed simultaneous sequencing of hundreds of samples[440].

While the development of large-scale dideoxy sequencing methods was progressing, another technology emerged; the Pyrosequencing. It uses a luminescent method for measuring pyrophosphate synthesis. ATP sulfurylase is used to convert pyrophosphate into ATP, which is then used as the substrate for luciferase, thus producing light proportional to the amount of pyrophosphate[441]. The sequence is inferred by measuring pyrophosphate production in a cycle, where each nucleotide is washed through the system over the template DNA affixed to a solid phase[442]. In each cycle, when the correct base is incorporated, as a consequence of pyrophosphate production, light is generated as the result of luciferin oxidation by luciferase (Figure 24)[443].

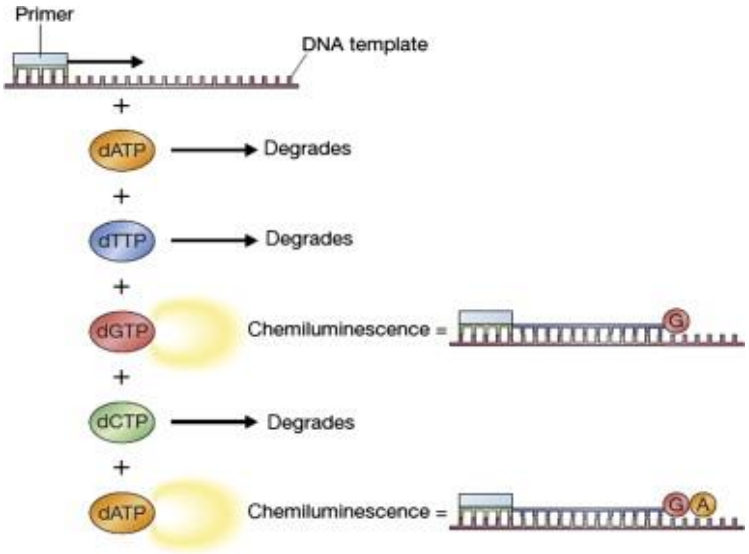


Figure 24: The Pyrosequencing chemistry. At each cycle, only when the correct base is incorporated, the reaction takes place, and pyrophosphate is produced. The light is generated as the result of luciferin oxidation by luciferase[443]

Like Sanger sequencing, also Pyrosequencing is a 'sequence-by-synthesis' (SBS) method, as the template is used to synthesize a complementary strand allowing the sequence to be inferred[443].

One advantage of the Pyrosequencing is the possibility to obtain the sequence in real-time, while the principal difficulty is to sequence homopolymer regions correctly[444].

Pyrosequencing was later licensed to 454 Life Sciences, becoming the first successful commercial next-generation sequencing (NGS) technology[425].

The 454 sequencing machine (after purchased by Roche), by mass parallelization of the sequencing process, increased consistently the possibility to sequence large amounts of DNA[445].

The introduction of mass parallelization increased the yield of sequencing efforts, allowing the sequencing of entire genomes (including the human genome) in a short time[446,447].

The 454 chemistry is characterized by the bond of the DNA molecules on microbeads, followed by emulsion PCR producing clonal DNA populations. These DNA-coated beads are then washed over a picoliter reaction plate containing wells large enough to fit only one bead. Then dNTPs are

washed over the plate, and pyrosequencing reactions occur (Figure 25 a, c). This set up was capable of producing reads around 400–500 bp[445].

Then, many other parallel sequencing techniques have been developed, and one among the others is Solexa (later acquired by Illumina)[448].

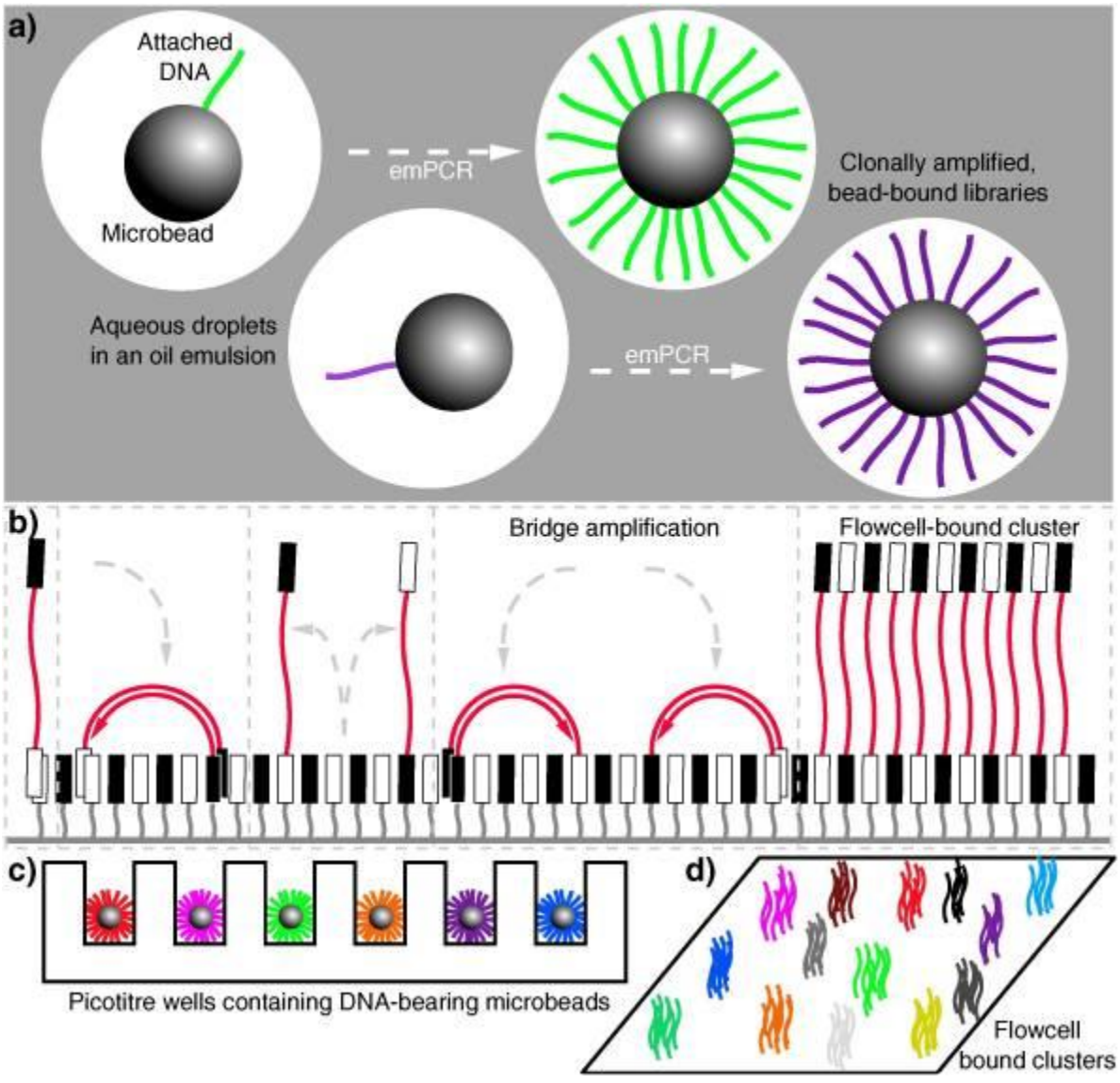


Figure 25: The parallelized amplification methods characterizing different second-generation sequencing technologies. (a): In 454 and Ion Torrent, the DNA binds to beads and is clonally amplified in an emulsion PCR and then sequenced. (b) Solexa or Illumina, are characterized by flow cells, where bridge amplification produces clusters of clonal DNA fragments in a planar solid-phase PCR, then sequencing occurs in an SBS manner. (c) Picotitre wells containing DNA-bearing microbeads (454 and Ion Torrent). (d) Flow cell clusters (Solexa and Illumina)[425]

In Solexa, the DNA fragments are bound to a flow cell trough complementary oligos fixed on the surface, and then “bridge amplification” generates clusters, representing the different populations of clonal DNA molecules. Then, SBS sequencing occurs using fluorescent ‘reversible-terminator’ dNTPs. In each cycle, one terminator with the fluorophore is bounded to the DNA molecule, the elongation stops, and the fluorescence is detected by a sensor. Four different fluorophores with different wavelengths are used to discriminate the four bases. In each cycle, the fluorescence of each of the sectors of the flow cell is recorded, and then these signals are converted in a DNA sequence by a software (Figure 25 b, d).

In 2005 Solexa presented the Genome Analyzer, capable of producing paired-end reads, where both the forward and the reverse of each sequence are generated. This machine was able to generate around one gigabase (Gb) of data per run[449]. These novelties together have been translated into the production of higher quality data. By 2014, the amount of data produced reached 1.8 terabases (Tb) in a single sequencing run.

Another remarkable second-generation sequencing platform is the Ion Torrent (Life Technologies). Similarly to 454 chemistry, DNA fragments are bound to beads, and clonal populations are generated through emulsion PCR and washed over a picowell plate. Then, in each sequencing cycle, the incorporation of a nucleotide causes the release of protons (H+), and the pH variation is detected thanks to the metal-oxide-semiconductor (CMOS) technology. This technology allows a high-speed sequencing, even sharing some weak points with the 454 technology like the difficulties with homopolymer sequeces[450–452].

The first human genome to be fully sequenced was co-published by Science and Nature and required 15 years of work and billions of dollars. In 2014, a new sequencing machine, the HiSeq X®Ten System, was placed in the market. This new technology was capable of sequencing 45 complete human genomes in parallel for the cost of 1000$ in a single day[453].

From the 1980s to today, the development of always newer technologies allowed the sequencing of ever longer DNA fragments, producing also increasing amounts of data (Figure 26) and thus expanding our science horizons[454].

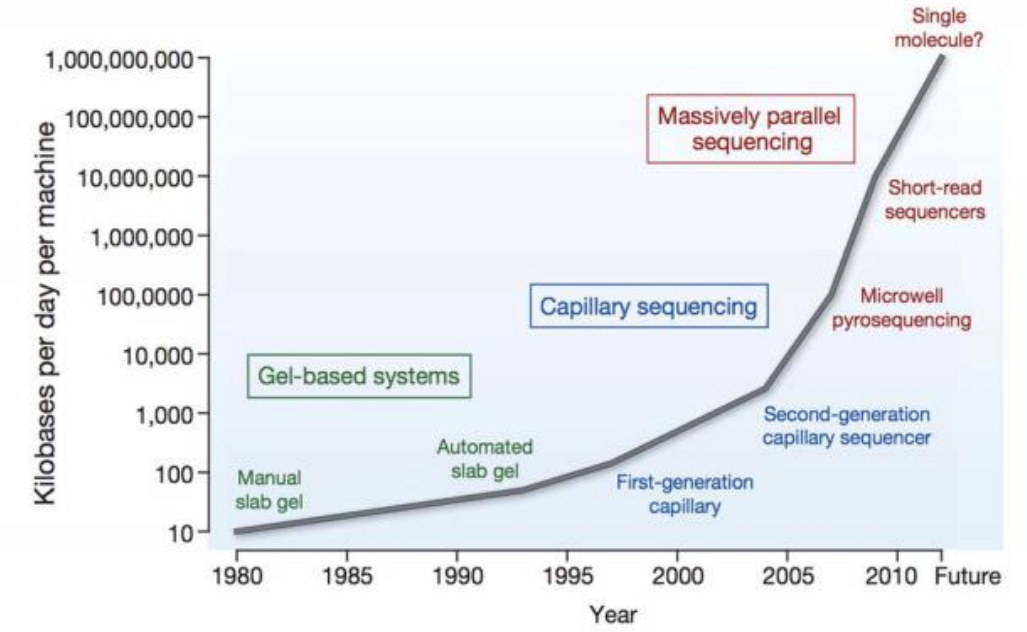


Figure 26: The evolution of the sequencing technology. In the x-axes, the year of presentation of the different new technologies. In the y-axes the amount (kilobases) of data produced in one day of a sequencing run by the various technologies[454].

5.3.1 NGS: Illumina sequencing technology

The advent of next-generation sequencing methodologies has revolutionized molecular biology. Some of the most successful NGS machines, like the Hiseq and the Miseq, are produced by the Illumina Inc company.

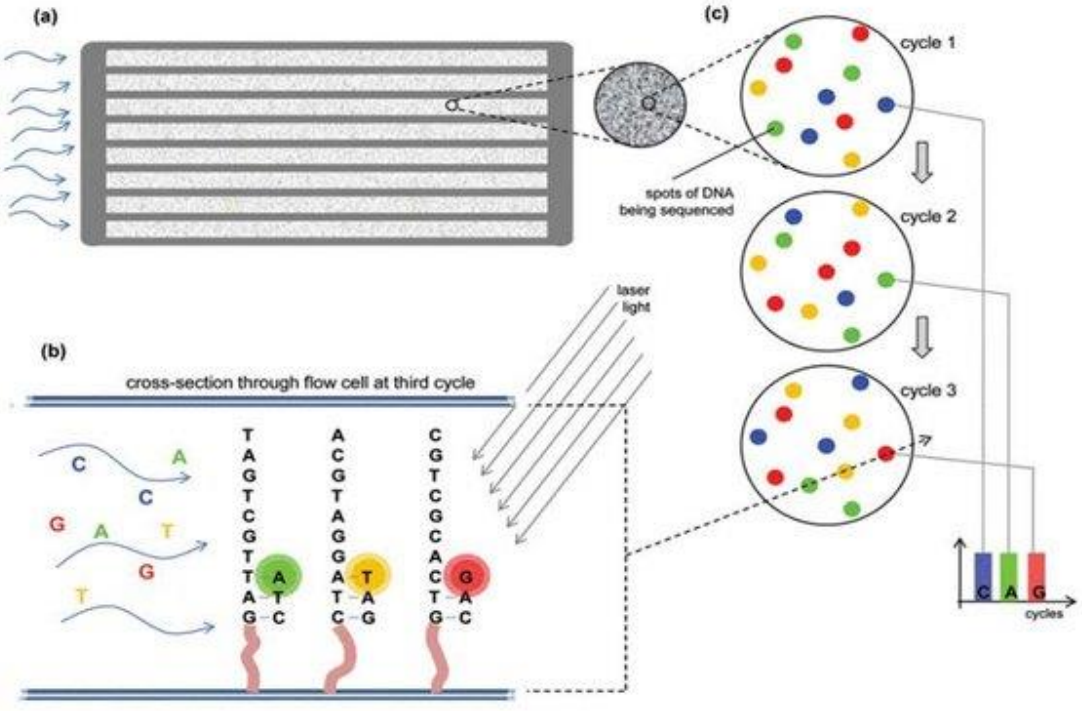


Figure 27: Sequencing by synthesis (Illumina). (a) Flow cell structure; (b) one nucleotide is incorporated in each cycle; (c) flow cell detail during the sequencing process, each incorporation is followed by emission of fluorescent light[455]

The principle of the Illumina sequencing is not so far from the previous technologies based on capillary electrophoresis. It is an SBS sequencing in which, in each cycle, fluorescently labelled deoxyribonucleotide triphosphates (dNTPs) are incorporated from the polymerase on the DNA template. Four different fluorophores are used to identify the four nucleotides, and in each cycle, the fluorescence is detected by a sensor (Figure 27). Mass parallelization is the feature that characterizes this methodology, together with the other high throughput sequencing methods, and billions of bases are sequenced in a single run. This technology generates very high quality reads with low error rates and a high percentage of base calls above Q30[456–458].

The so-called paired-end sequencing permits to sequence both the forward and the reverse strand of a DNA fragment represented by one cluster on the flow cell, thus allowing to obtain

higher quality sequences (Figure 28), especially when homopolymeric sequences are present, reducing the sequencing error rates in these regions[459,460].

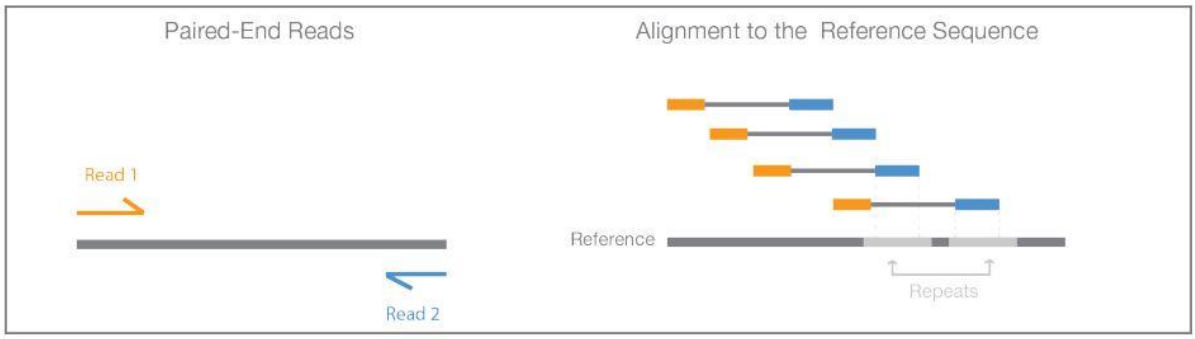


Figure 28: Illumina paired-end sequencing. Both the 3' and 5' of the DNA fragments are linked to adaptors, and both forward and reverse strands are sequenced. Then, through in silico analyses, it is possible to pair them obtaining higher quality consensus sequences. Millions of reads are produced in a sequencing run and can be aligned to a reference genome[454]

The first step of a typical Illumina sequencing protocol is the library preparation, where the DNA template to be sequenced, is fragmented through a tagmentation process. Then Illumina adapters containing either one (i7 index) or two index sequences (i7 and i5 index) for each sample library are added (Figure 29 A). Different libraries can now be pooled (multiplexed), sequenced on the same flow cell, and de-multiplexed with later in silico analyses, allowing cost savings and experiment scalability[461]. Follows the cluster amplification step, where the DNA fragments are loaded on the flow cell, where DNA fragments hybridize with the oligonucleotides fixed on the solid surface. After bridge amplification occurs, and the different DNA fragments are amplified, generating clonal clusters (Figure 29 B). Now the SBS sequencing can start. At each cycle, one base is added, and the fluorescence produced by each of the clusters is recorded. In each cycle, just one nucleotide can be added because a fluorophore occupies the 3′ hydroxyl position on the just bonded nucleotide. This must be cleaved away before polymerization can continue, which allows the sequencing to occur in a synchronous manner[448,462] (Figure 29 C).

The last step of this process is the data analysis (Figure 29 D). The machine generates paired FASTQ files where forward and reverses reads can be identified through their ID. Additional steps can be performed to align the reads to a reference genome or to assemble them for the generation of longer consensus sequences. Different kinds of analyses can be performed, such as single nucleotide polymorphism (SNP), insertion-deletion (indel) identification, transcriptome, phylogenetic or metagenomics, and much more.

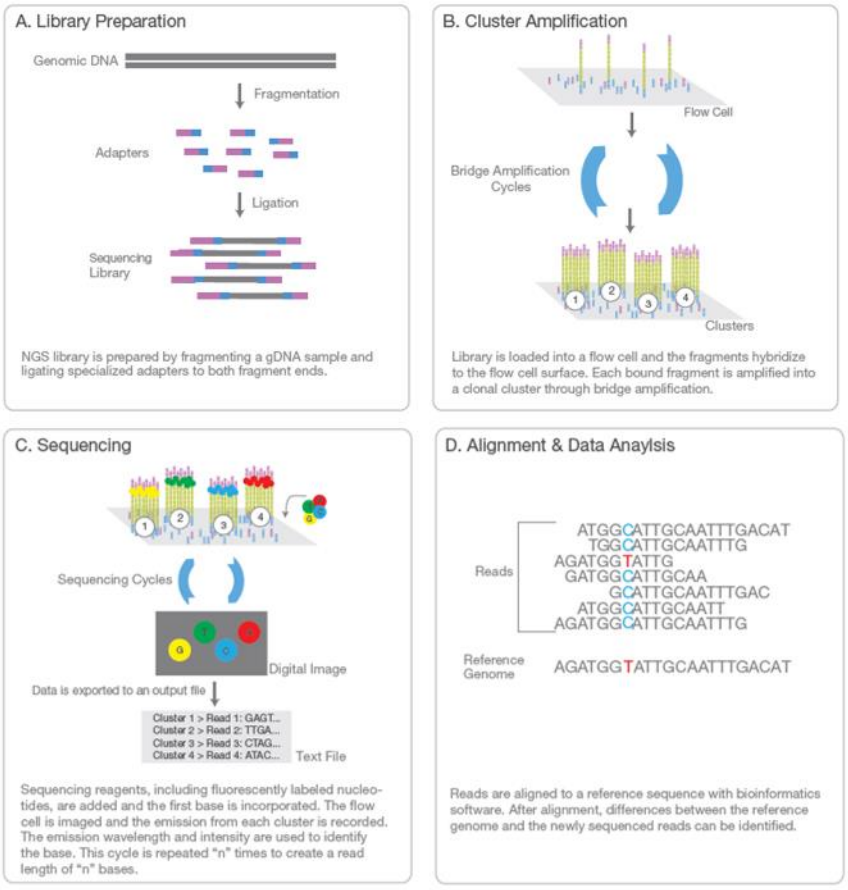


Figure 29: Illumina NGS workflow. (A) Library preparation. (B) Cluster amplification. (C) SBS sequencing. (D) In silico data analysis. A quality score is associated with each of the bases recognized from the machine, and quality trimming is performed to obtain high-quality sequences[454].

One of the first things to do in the data analysis is the raw reads quality check. The reads quality can be assessed using tools like FASTQC. Figure 30 shows a typical bases quality score graph

generated using FASTQC. A decrease in base quality with the increase of the base position is quite common (Figure 30). The SBS chemistry underlies the occurrence of this phenomenon.

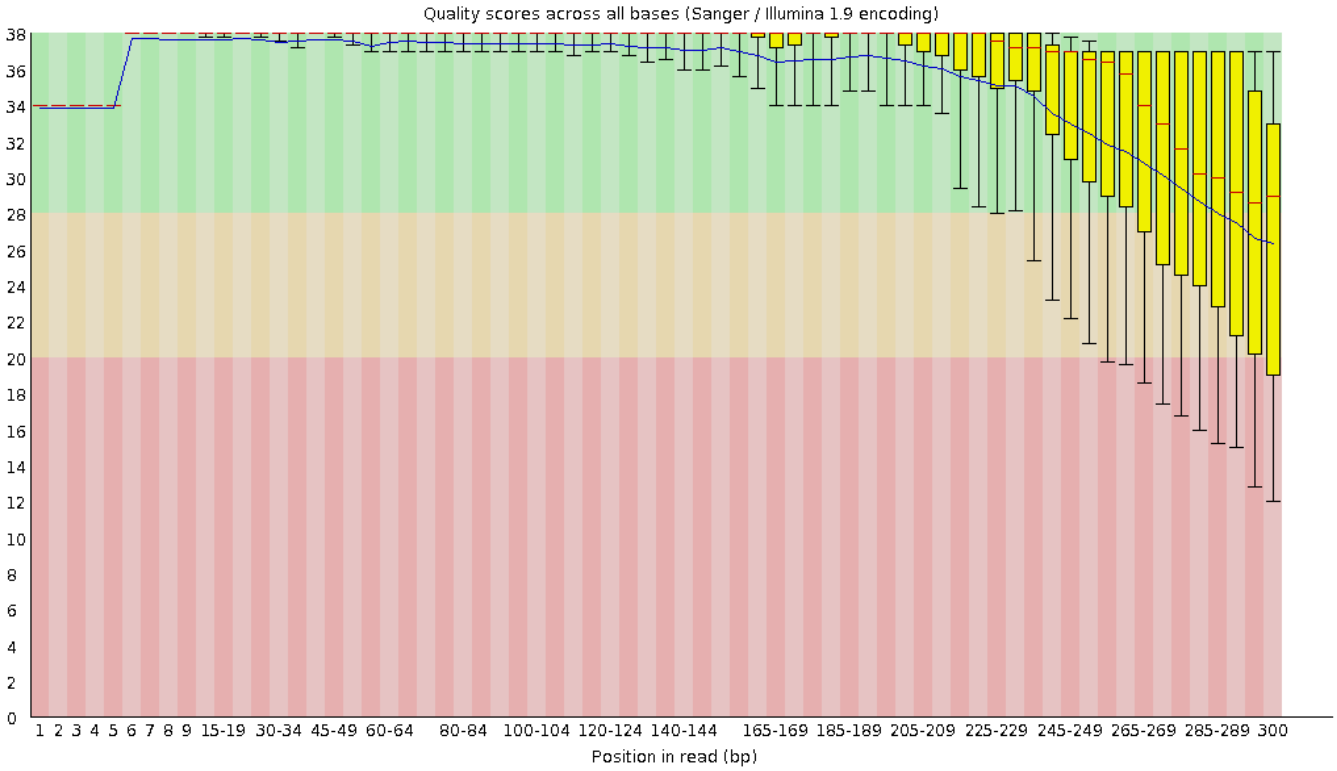


Figure 30: FASTQC quality control with a typical decrease of the quality over the read[463]

In each sequencing cycle, a nucleotide is added to the sequence. A fluorophore is present in the 3' (terminator), and no other nucleotides can be added in the same cycle. If the terminator is not correctly removed, in the next cycles, that specific DNA fragment will be n-1 shorter compared to the other sequences of the same cluster. In the next cycles, this DNA fragment will produce a different fluorescent signal, compared to the other sequences of the same cluster. In this context, a lower quality will be registered by the detector, since the quality is computed based on the intensity of the fluorescent signal. This phenomenon is named phasing (Figure 31 Left)[464]. Another similar phenomenon is the prephasing, where, in a single cycle, more than one nucleotide is

added because of a defective terminator (Figure 31 Right)[464]. In this case, from the next cycle, the sequence will have n+1 nucleotides compared to the other sequences of the same cluster.

In the first steps of the sequencing, these kinds of errors occur with a low probability. However, with sequencing progression, they tend to be accumulated, and the signal gets more and more asynchronous, affecting the quality of the bases in the last part of the fragments. A workaround, used by the companies, is to develop chemistries with more cycles, helping in the sequencing of longer fragments before the quality drop[463].

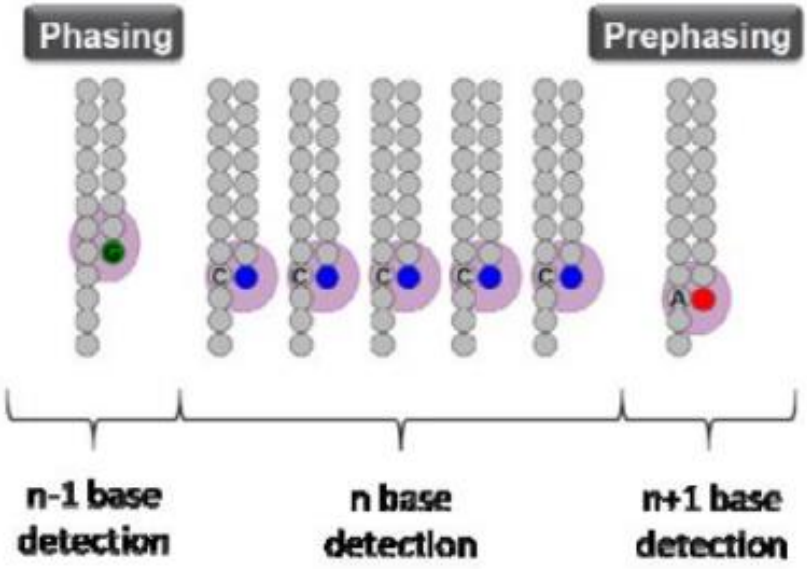


Figure 31: Phasing (Left) and prephasing (Right) processes that can affect the base quality, mainly at the end of the reads[464]

5.3.1.1 Targeted Sequencing

Some different strategies and protocols can be applied to best fit with the needs of a particular experimental condition and to maximize the yield. For example, targeted sequencing strategies are particularly useful when the objective is to amplify a specific region of a genome or to increase

the coverage of a DNA sequence to be reconstructed, especially when the target sequence is particularly long and if it is in low copy number.

A specific DNA region can be amplified using a targeted sequencing strategy, and after the sequencing, being represented by hundreds or even thousands of times. This strategy is particularly useful in low-frequency genetic variants studies.

Commercial NGS sequencing panels for specific analysis are available, including kits with probe sets focused on particular areas of interest such as cancer, cardiomyopathy, or autism. Moreover, custom targeted sequencing protocols can be developed. Custom probe sets can be produced, enabling researchers to target regions of the genome relevant for specific studies.

Some other applications of the targeted sequencing are studies on genes in specific pathways, follow-up in whole-genome sequencing (WGS) studies, and genome-wide association studies (GWAS).

There are two main targeted sequencing methods proposed by Illumina. The first, named “target enrichment”, is characterized by the binding of the target DNA with biotinylated probes and then purified by magnetic pulldown (Figure 32 A). The second one, named “amplicon generation”, involves the amplification and purification of the target DNA using highly multiplexed PCR protocols. In this strategy, first custom primers are used to amplify the specific region of interest. Then the sequencing adaptors are added to the 3’ and 5’ of the amplicons through another amplification step or ligation (Figure 32 B). In the final product, the sequence copy number will be significantly enriched. This second approach is particularly useful in genetic variant studies, reconstruction of whole bacterial or viral genomes, the discovery of rare somatic mutations in complex samples, such as cancerous tumors mixed with germline DNA, bacterial 16S rRNA gene analyses in taxonomy or metagenomics studies[465–467].

More recently, amplicon sequencing has been used for HPV genotyping, and it showed to be very useful in the discovering of new HPVs[468–470].

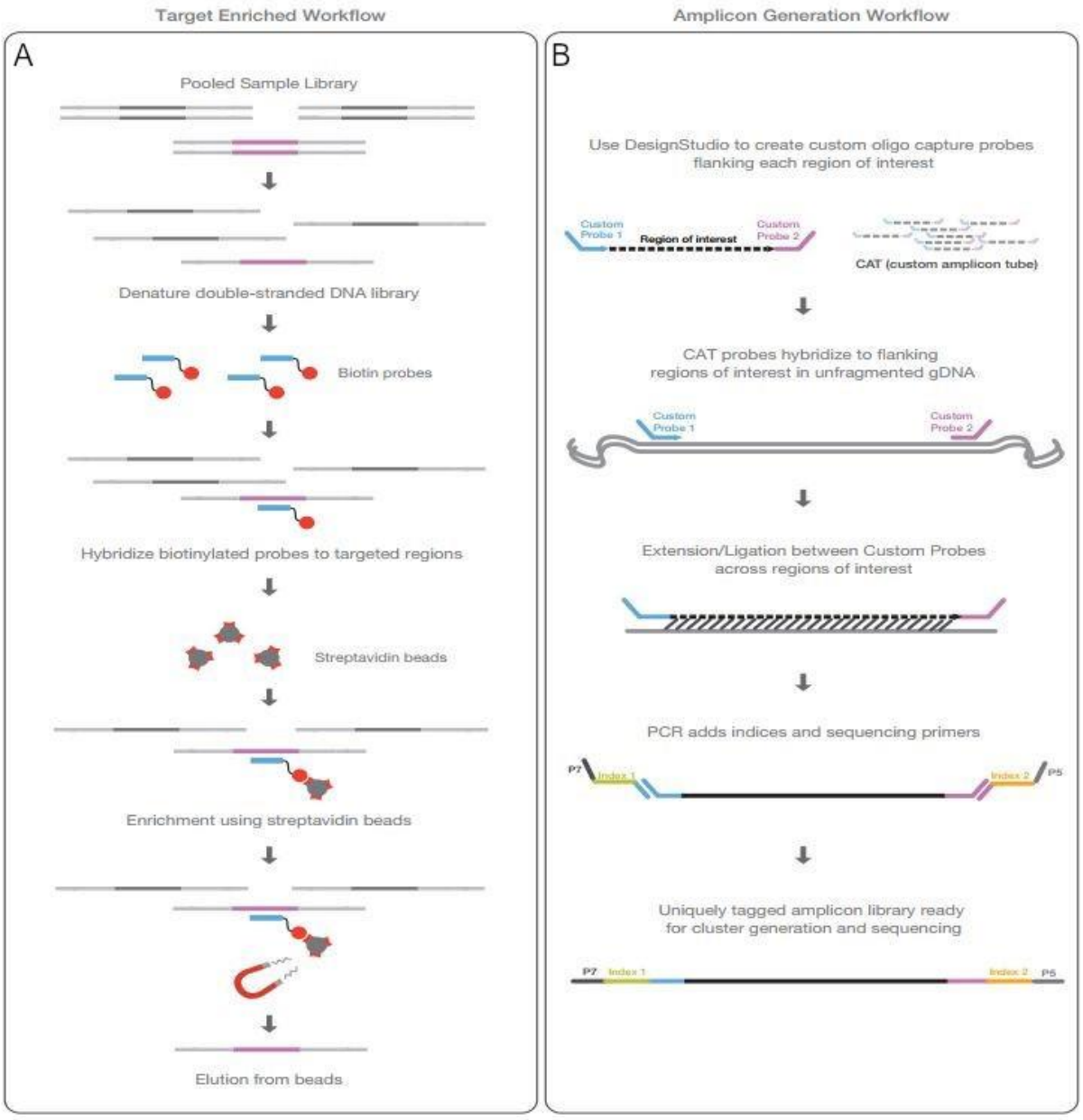


Figure 32: Targeted sequencing workflow. (A) Target enriched workflow; specific regions of interest are hybridized to biotinylated probes and then purified by magnetic pulldown. (B) Amplicon generation; Specific regions of interest are amplified using highly multiplexed PCR primers, and then sequencing adaptors are added to the DNA fragments[454]

5.3.2 RCA and NGS for HPVs identification

Next-generation sequencing technologies are successfully applied to the characterization of Human Papillomaviruses (HPVs). In this context, modern NGS methodologies showed their worth both in the characterization of the HPV virome in human samples and the identification of new viruses[470–473]. Studies on HPV integration sites, disrupted genes and pathways, and common and distinct genetic and epigenetic alterations, are just some of the fields where NGS is giving his substantial contribution to understanding the role of the different HPVs in human diseases[474,475].

In 2017, the effectiveness of the HPV genotyping performed using NGS methodologies and previously developed and standardized *in-situ* hybridization protocols were compared. NGS has shown a sensibility 22% higher if compared to the hybridization method and also a higher resolution allowing the identification of genetic variants[476].

In a metagenomics study published in 2018, Pastrana and colleagues discovered 83 novel HPVs, in immunocompromised patients affected by different pathologies (warts, hypogammaglobulinemia, infections, myelokathexis syndrome, or EV). In this study, virion enrichment (RCA) was used in combination with NGS[477].

Contigs were generated from the raw reads, with a *de novo* assembling approach (SPAdes)[478]. After, the contigs were identified using Megablast algorithm against the whole nr/nt NCBI database. Further analyses allowed the identification of known and putative new HPVs, and the reconstructed sequences were confirmed by Sanger sequencing[477].

A phylogenetic tree was generated, representing the 83 novel HPVs detected (Figure 33).

From the 83 new HPV types, 69 were Gamma types, 8 are Beta types, 1 is Mu, 1 is Alpha, and 4 were potential new HPV species (<70% identity). Incomplete genomic sequences of an additional 35 potentially novel Gamma types (including some possible new species) were also observed. No representatives of new papillomavirus genera (L1 with 60% identity to the nearest neighbor) were found[477].

A potential pitfall of massively parallel sequencing is that contig assembly can result in the artifactual assembly of chimeric genomes, mainly when closely related viral strains are present in the same sample[477].

In the work of Pastrana and colleagues, to check the presence of chimeric sequences, phylogenetic trees were generated, based on E1 and L1 protein sequences, followed by evaluation of the topology for the novel sequences[477].

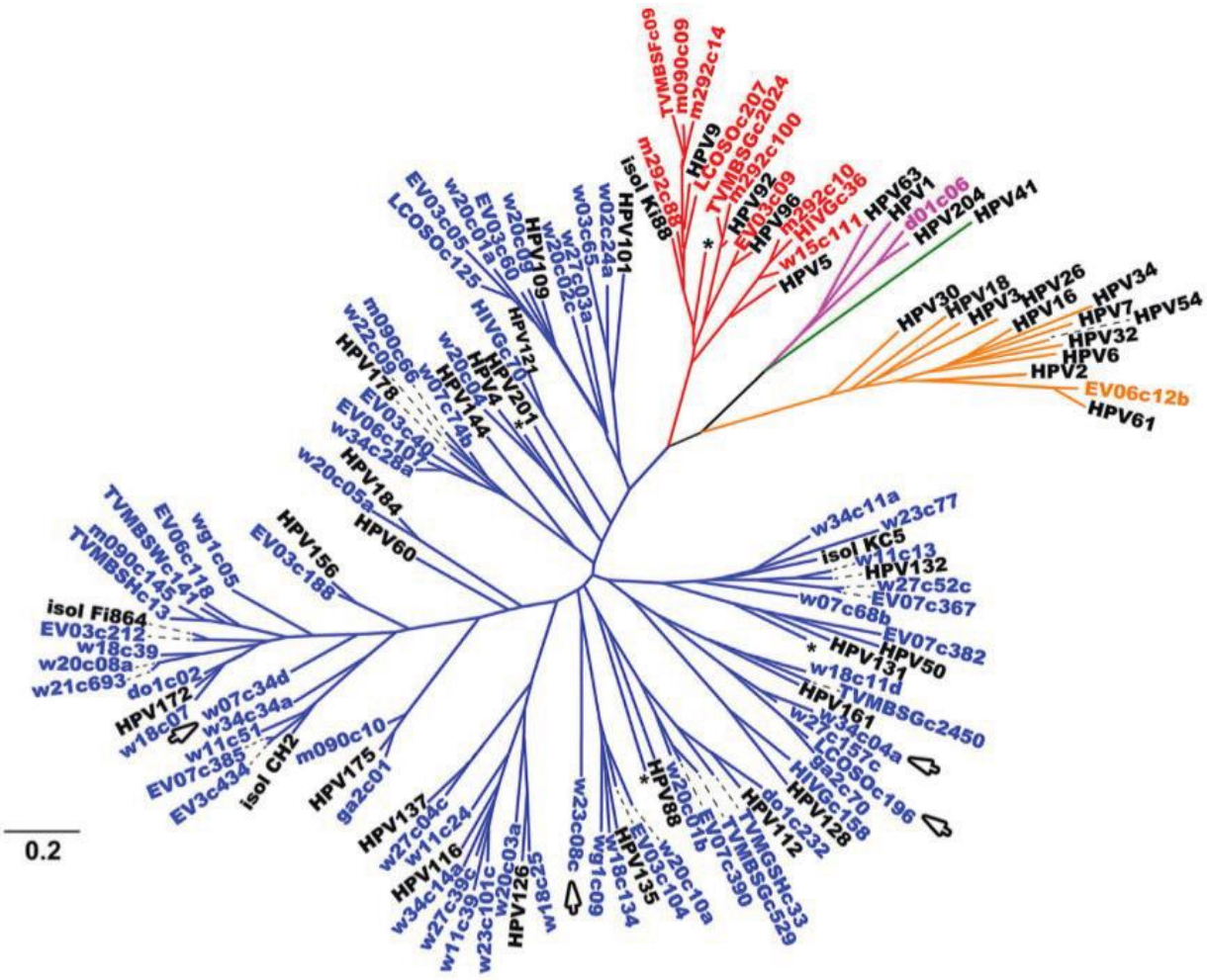


Figure 33: Phylogenetic tree representative of the 83 novel HPV types discovered, based on the L1 protein sequence identity. At least one HPV type from each previously known species (black font) was analyzed, along with each of the 83 complete HPV genomes catalogued. Asterisks show known representative species that did not fit due to the figure's space restrictions (counterclockwise, Beta 3-HPV49, Gamma 20-HPV163, Gamma 21-HPV167, and Gamma 2-HPV48). HPV genera are shown with different colors: orange, Alpha; red, Beta; blue, Gamma; pink, Mu; green, Nu. Arrows indicate potential new species[477]

In another work published in 2018 by Tirosh and colleagues, 250 putative new HPV genomes were identified from dedicator of cytokinesis 8 (DOCK8)-deficient skin samples. DOCK8-deficiency is a rare primary human immunodeficiency characterized by recurrent cutaneous and systemic infections, as well as atopy and cancer susceptibility[479].

Contigs were obtained from the raw data, using a *de novo* assembling tool (SPAdes)[478] and the sequences were identified based on the BLASTN best hit against the whole nr/nt NCBI database[480]. Using the L1 taxonomy tool available on the PaVe website (https://pave.niaid.nih.gov/#analyze/l1_taxonomy_tool), 205 out of the 250 novel genomes were depicted as novel types, and 45 genomes were depicted as members of a single novel species. These novel species were subsequently verified by targeted PCR to the L1 region from the original DNA extracted from the skin swabs. In all, 229 of the 250 novel HPV genomes belonged to the gamma genus of HPV, 19 belonged to the beta genus, and two belonged to the mu genus[481].

5.3.3 PCR-based and WGA NGS for HPVs identification

The use of PCR primers for the amplification of HPV sequences in combination with NGS is a valid strategy for the study of the HPV diversity in human samples[209,400,401].

In 2011 Ekström and colleagues used the FAP primers to amplify HPV sequences in biopsies from different skin lesions (e.g., squamous cell carcinoma, actinic keratoses, and keratoacanthomas), followed by high throughput sequencing[400]. This work allowed the identification of 44 putative novel HPV types. The raw reads were processed to assemble contiguous sequences and then identified using BLASTN against the GenBank database (NCBI). Known and putative new HPV types were discriminated considering the percentage of identity with the closest known HPV types. In addition to the identification of many known HPV types in the different samples, 44 novel putative HPV types were identified, most of which belonging to the gamma genus[400].

In 2013, Ekström and colleagues published another work where different skin samples, representative of different pathologies (i.e., squamous cell carcinoma, actinic keratosis, and keratoacanthoma), were analysed for the detection of known and putative new HPV types. An extended HPV general primer PCR and high-throughput sequencing of amplimers were used in this analysis. In this study, FAP59/64 primers were used alone and in combination with five forward and four reverse partially degenerate new primers.

The raw data were analysed to remove chimeric sequences, assembled using MIRA software[482], and identified using Blastn against GenBank database (NCBI)[483].

This work allowed the identification of 273 different HPV isolates (87 known HPV types, 139 previously known HPV sequences (putative types), and 47 sequences from novel putative HPV types). Among the new sequences, five were representative of the genus beta and 42 of the genus gamma. The Luminex assay was also used to confirm the presence of the putative new HPVs, and only a part of them was identified in various skin samples using this technique, probably because of low copy numbers or mismatches between the probes used and the viral sequences.

In 2015, the same research group used another strategy to detect HPV sequences from different lesions (squamous cell skin carcinomas, keratoacanthomas, actinic keratoses, basal cell carcinomas, and SCCs *in situ*)[209]. Whole-genome amplification (WGA) was performed using Illustra™ Ready-To-Go™ GenomiPhi™ DNA Amplification Kit (GE Health Care, United Kingdom), and the sequencing was performed without any specific PCR amplification step[209].

For comparison, a pool of the same samples after general primer PCR amplification was also sequenced. Not all the HPVs were detected using this second approach, indicating that sequencing without prior PCR gives a more unbiased representation of the HPVs present. On the other hand, the use of a PCR-based strategy can help in the detection of HPV sequences less represented. This work allowed the identification of 10 different HPVs in 47/91 specimens. These sequences represented four established HPV types (HPV types 16, 22, 120, 124), two previously

known putative types (present in GenBank), and four previously unknown HPV sequences (new putative types). The most commonly detected virus was cloned, sequenced, and designated as HPV197[209].

To conclude, both WGA and PCR-based approaches are useful in the identification of known and new HPV types, but with different specificity and sensitivity[209,401].

5.3.4 Third-generation sequencing

The evolution of DNA sequencing technologies has always promoted scientific discoveries. In particular, the advent of high-throughput technologies has driven the rapid progress of life science[484]. The last milestone in DNA sequencing technologies is the advent of third-generation sequencing.

Third-generation sequencing technologies are capable of analysing long DNA molecules and don't require a DNA amplification step, shared by all the previous technologies.

The first third-generation sequencing technology was developed by Stephen Quake in 2003 and is based on a single molecule sequencing methodology (SMS)[485,486].

It works similarly to Illumina technology but without bridge amplification. The DNA fragments are bound on a solid surface, and then fluorescent terminator nucleotides are used in the sequencing process. Cycle by cycle, the sequence is elongated, and a fluorescent detector record the signal then converted in nucleotide sequence. The absence of clonal amplification in this technology reduces the sequencing error rate[487,488].

One of the most used third-generation technologies is the PacBio from Pacific Biosciences[489], a single-molecule real-time (SMRT) platform performing the DNA polymerization on nanostructures called zero-mode waveguides (ZMWs), which are tiny holes in a metallic film covering a chip. The properties of these wells allow to generate a light signal only at the very bottom of the well, where the new nucleotide is incorporated, and the signal is recorded from the chip (Figure 34 a). The

ZMWs structure allows the generation of a very clean light signal that makes this technology effective in sequencing[490]. The sequencing can be monitored in real-time, and in each well, a single polymerase with a single DNA molecule is present.

This process can sequence single molecules in a small amount of time. Some of the most interesting features of this technology are the possibility to identify modified bases and the considerable length of DNA fragments that can be sequenced (up to 10 kb)[487,489].

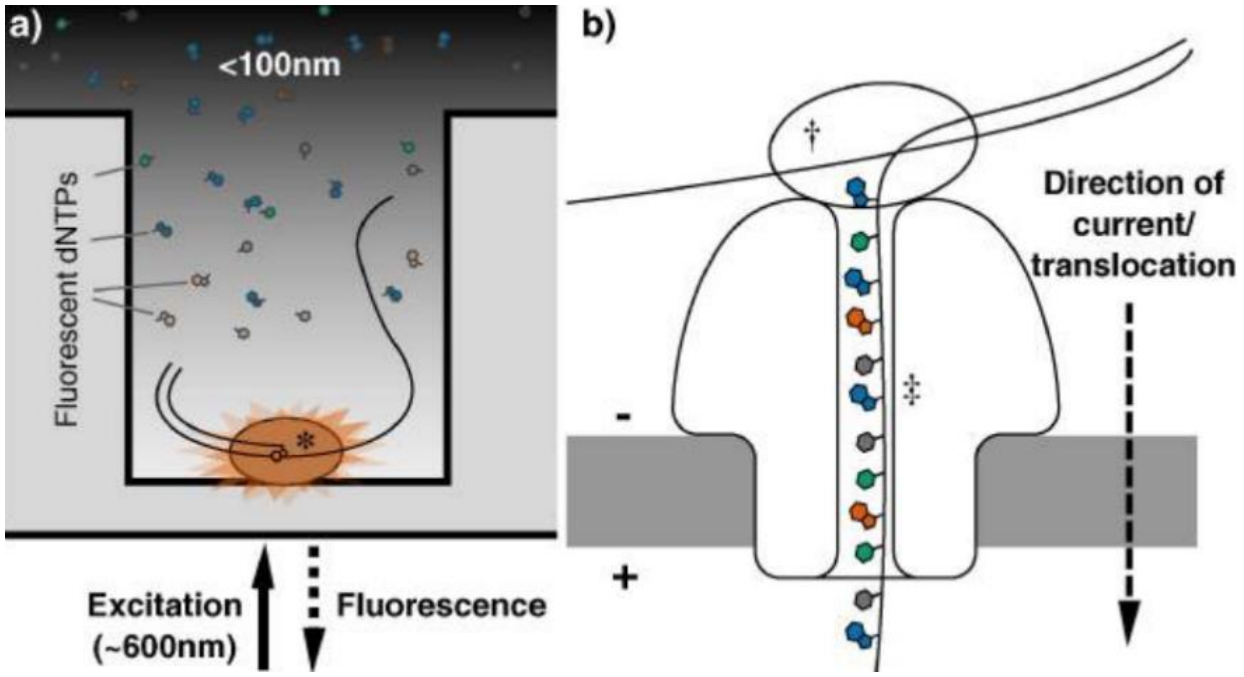


Figure 34: Third generation sequencing technologies. a) PacBio, zero-mode waveguides (ZMWs) technology. b) MinION Nanopore sequencing technology[425]

Another third-generation sequencing technology was developed later in 2014, from Oxford Nanopore Technologies (ONT). They presented two sequencing machines, GridION and MinION capable of sequencing very long DNA fragments using the nanopore technology, developed years before (Figure 34 b)[491,425,460,492,493].

All the previous sequencing technologies are based on sequencing by synthesis (SBS) technology. The nanopore sequencing drives the single helix of a DNA molecule through a

nanopore, and the sequencing takes place by the detection of ionic current changes, without DNA synthesis processes[494,495].

In nanopore sequencing, a pore-forming protein is set in an electrically resistant polymer membrane. The nanopore works as a biosensor, representing the unique connection between two sides of the membrane. A constant voltage bias produces an ionic current and drives ssDNA through the pore. An enzyme (e.g., a polymerase or helicase) named “motor” is bound to the DNA and helps the opening of the double helix and the passage through the nanopore, nucleobase by nucleobase. An ionic current is passed through the nanopore by setting a voltage across this membrane. The ionic conductivity through the lower part of the nanopore is particularly sensitive to the nucleotides and their specific mass. When a nucleotide passes through the pore, this event creates a characteristic disruption in the current. Each nucleotide generates a specific current alteration that is translated in his identification (Figure 35)[496]. In the newest chemistries, the stepping rate is more than 300 bases per second[497].

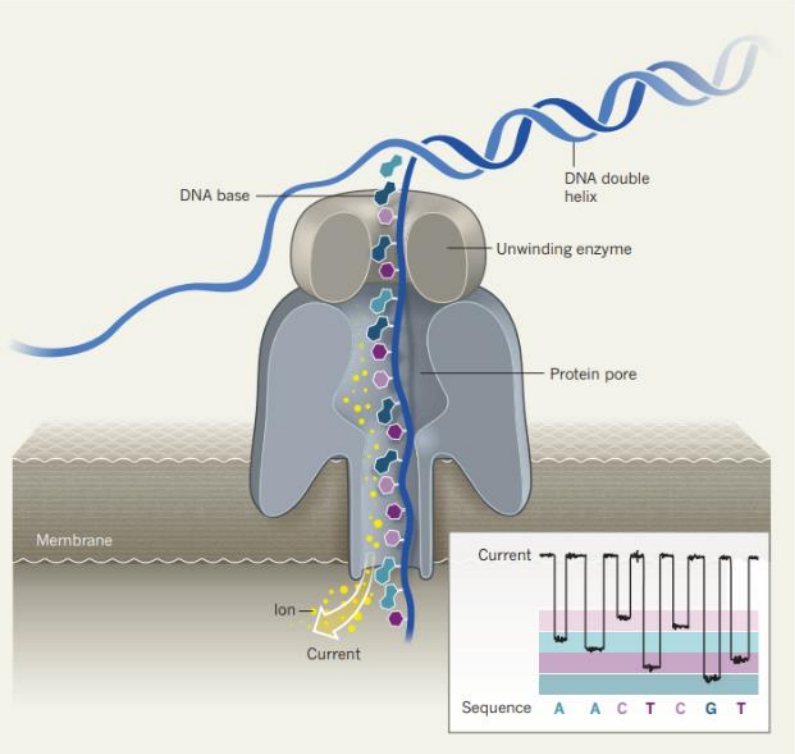


Figure 35: Detail of nanopore technology. Two chambers filled with ionic solutions are separated by a voltage-biased membrane. An unwinding enzyme (helicase) drives a single helix of the DNA through the nanopore. Ions and DNA molecules pass through the nanopore from one side to the other of the membrane. Variations of the ionic current are measured by a sensitive ammeter and thus converted into nucleotide sequence[498]

According to the kind of sample that has to be processed, there are different strategies to optimize the protocols, control the read length, or maximize the throughput.

During the library preparation, both the template and the complement strand are linked with adaptors and motor proteins. Thus, both the forward and the reverse strand of a DNA fragment can be sequenced (Figure 36)[499].

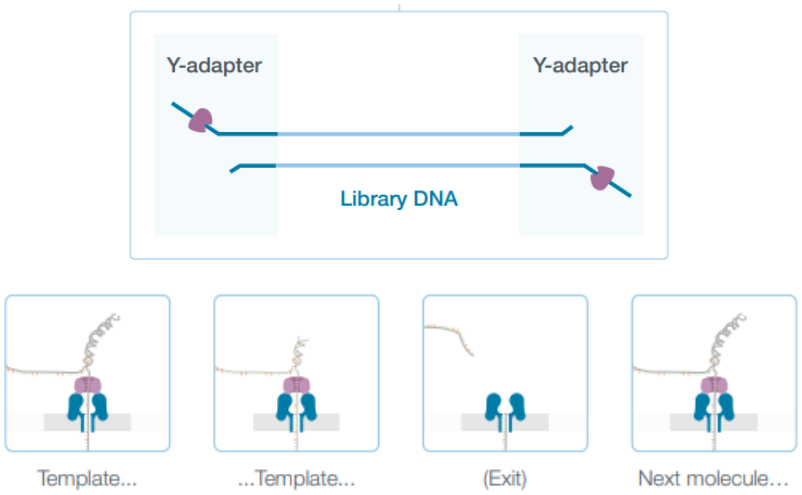


Figure 36: Nanopore sequencing. Both the template and the complementary strand can be carried into the nanopore and sequenced. During the library preparation, sequencing adaptors and motor proteins are attached to the ends of the double strand[499]

The sequencing adaptors can be added to the DNA template through a ligation step (Figure 37 Left), after end-prep and nick repair or through transposases activity for a faster protocol (Figure 37 Right)[499].

This technology can be used to answer different scientific questions: whole-genome sequencing, targeted sequencing, metagenomics, epigenetics, and even direct RNA sequencing. The most significant advantages of this technology are the portability and the rapidity of the library preparation. These characteristics make this technology particularly useful, for example, in experiments on the field[499].

The main problem of this technology is the high error rate in the sequencing. Nevertheless, it represents a new cost-effective technology being quite cheaper compared to the previous machines, suitable for the sequencing of very long DNA fragments, such as viral and bacterial genomes[494,500]. Moreover, the portability of this technology also allows the sequencing in limiting conditions, such as when instant results are required on the field[491].

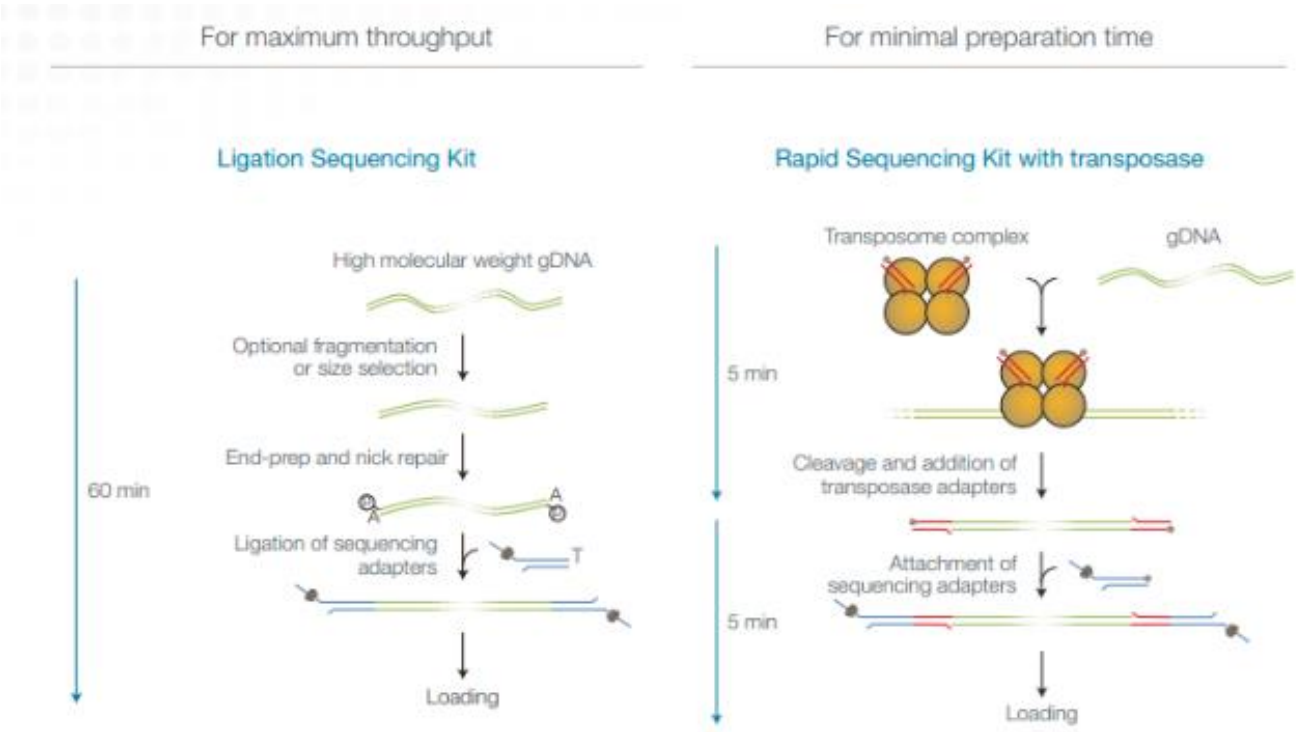


Figure 37: DNA library preparation protocols for nanopore sequencing. (Left) Ligation sequencing protocol. (Right) Rapid sequencing protocol with transposase[499]

5.3.4.1 Oxford Nanopore Technology for the sequencing of papillomaviruses

Third-generation sequencing methodologies can be applied in the identification and characterization of papillomaviruses. The possibility to process longer sequences, compared to the previous technologies (i.e., Sanger and NGS), can lead to a faster and efficient reconstruction of the viral genomes.

In a study conducted by Vanmechelen and colleagues in 2016, the Nanopore sequencing technology was used for the identification of novel species of papillomavirus (PV), in warts of giraffes (*Giraffa camelopardalis*)[501].

Wart tissue was excised from the giraffe skin biopsies, and amplification of the viral genomes was performed using rolling circle amplification (RCA)[501]. The RCA product was digested with EcoRI

and BamHI to reduce the presence of potentially co-amplified plasmids. The resulting fragments were separated on an agarose gel (0.9%) to look for bands that had a cumulative size of ~8 kb. These specific bands were extracted and purified. Extracted DNA fragments were then subjected to multiple displacement amplification (MDA) using the REPLI-g Mini Kit (QIAGEN)[501].

Both digested fragments and undigested RCA products were pooled and sequenced using MinION sequencer. The sequencing library was loaded onto a primed MinION flow cell (R7.3) and sequenced for 22 and 34 hours for runs 1 and 2, respectively. Primer walking Sanger sequencing was conducted to confirm the sequences obtained with the MinION sequencing[501].

Different tools were used to process the sequencing data. Metrichor (v2.39.3; https://metrichor.com) was used for base calling. Poretools (Loman and Quinlan, 2014) was used to retrieve the sequence data in FASTA format from the FAST5 files generated by Metrichor. Tblastx software (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to identify the reads that displayed similarity to deltapapillomaviruses. Canu[502] was used to assemble the identified viral sequences, and Nanopolish[500] was used to correct the resulting contig further. The NCBI ORF finder and Blast[483] tools were used to detect ORFs and identify them based on their similarity with other Deltapapillomavirus ORFs[501].

This protocol allowed the complete characterization of a new deltapapillomavirus named Giraffa camelopardalis papillomavirus 1[501].

In another work published in 2019 by Quan L. and colleagues, cervical cancer samples were analysed both for HPV infections and microbiome[503].

In this work, multiplex PCR, targeting both HPV16 E6/E7 and full-length 16S rRNA, and Nanopore sequencing were combined.

The Nanopore sequencing results were compared with Illumina sequencing results and showed a similar microbiome composition[503].

QUIIME and LAST were used to perform the taxonomic classification of the sequences. Different databases were used to obtain better annotation results: the HPV genome, the NCBI 16S ribosomal RNA (Bacteria and Archaea) database, and the GreenGenes database[503].

A minimum of 15 min Nanopore sequencing was enough to identify the top 10 most abundant bacteria. Furthermore, two HPV integration sites were identified and verified by Sanger sequencing. This method has many potential applications in pathogen identification and can potentially help in providing a more rapid diagnosis[503].

Both these studies confirm the effectiveness of the MinION technology in the identification and characterization of novel papillomaviruses.

5.4 The sequencing data analysis

Nowadays, the use of computers and specific software has become an essential part of the biologist's work. In modern science, there is often a need to parse meaningful information from large sets of data, and this became even more clear with the advent of high-throughput sequencing technologies that changed all the scientific approaches[504].

Margaret Dayhoff (1925–1983) was an American physical chemist who pioneered the application of computational methods to the field of biochemistry and is considered the mother of the bioinformatics[505]. She developed a program in FORTRAN for *de novo* protein sequence assembling. Dayhoff and Eck in 1965 created the first-ever biological sequence database representing amino acids with single letters. This first protein database helped in developing the early theories about the evolutionary relations between proteins from different species and introducing the new term "Paleogenetics": the study of the evolution of the species based on the proteins sequence similarity[506].

In 1970 Needleman and Wunsch[507] developed the first dynamic programming algorithm for pairwise protein sequence alignments.

In the 1980s, the first multiple sequence alignment (MSA) algorithms emerged. This tool is based on a generalization of the Needleman–Wunsch algorithm, which involved the use of a scoring matrix whose dimensionality equals the number of sequences[508]. This first approach was not very efficient because of the long time required to find the correct alignment.

Da-Fei Feng and Russell F. Doolittle, in 1987, developed the first valid algorithm for MSA[509]. Also, their tool uses the Needleman–Wunsch algorithm. From each pairwise alignment, it extracts pairwise similarity scores and uses these scores to build a tree. Then, the tree guides the alignments between the most similar sequences[504].

In 1988 the famous CLUSTAL software was released. It was based on the Feng–Doolittle algorithm[510] and is still maintained and used to the present day[511].

While the evolution of protein sequences alignment tools was proceeding, several milestones in molecular biology were setting DNA as the primary source of biological information. In 1977 Sanger developed the first effective DNA sequencing method, and the principle of this technology is the basis of the modern Sanger-based machines[512].

Roger Staden, in 1979, published the first software suite for the Sanger sequencing data analysis[434]. This suite could manipulate DNA sequences, search overlaps, and generate contigs[504].

With the improvements of the DNA sequencing technologies, larger quantities of genetic information became available. It was immediately apparent that the DNA is much more informative because it can contain the trace of an evolutionary event not visible at the protein level[504].

The study of DNA sequences has led to the development of new algorithms and models to represent the evolutionary relations between different organisms.

The maximum parsimony method is one of the main algorithms used to generate evolutionary trees. It is based on the assumption of the maximum parsimony as the primary mechanism driving evolutionary changes.

Another notable method for the inference of phylogeny from DNA sequences is the maximum likelihood (ML) method. ML tests which trees yield the highest probability of representing the observed data[513].

Nowadays, many models for the inference of phylogeny are based on the maximum likelihood (ML) method.

In the 1990s, Bayesian statistics was also introduced in the molecular phylogeny and is still widely used[514,515].

The first programming languages used to develop bioinformatics tools were difficult to implement and not very intuitive. Thus the approach of biologists to these languages has not been immediate.

The progressive evolution of the programming languages has led to the generation of even more intuitive and easy-to-use programming languages, facilitating the development of novel algorithms for the analysis of biological data.

Nowadays, the two major programming languages used by most of the bioinformaticians are Perl and Python.

Perl (Practical Extraction and Reporting Language), was introduced by Larry Wall in 1987 as an addition to the GNU operating system to facilitate parsing and reporting of text data[516]. Until the late 2000s, Perl was the principal programming language used by most bioinformaticians, because of its great flexibility[517] and the development of BioPerl in 1996 contributed to his

popularity in life science[518]. BioPerl provides several modules to facilitate the work of a bioinformatician. It can help in accessing sequence data from local and remote databases, switching between different file formats, similarity searches, and annotating sequence data.

The implementation of Python started in 1989. Python is characterized by a more straightforward vocabulary and syntax to make code reading and maintenance simpler[504]. After the 2000s, Python becomes the dominant programming language used in bioinformatics[519].

The information from other researchers and their progress became gradually more accessible, and several public databases appeared. In 1992 the NCBI made available the GenBank. In the same period, the EMBL Nucleotide Sequence Data Library, including databases such as SWISS-PROT and REBASE, was released[520,521].

In 1994 the BLAST tool appeared in the NCBI site, together with other famous databases: Genomes (1995), PubMed (1997), and Human Genome (1999)[504].

Always newer sequencing technologies allowed the production of much more data, resulting in an exponential increase of sequences in public databases such as GenBank and WGS and further preoccupations towards Big Data issues. The scientific community has now generated data beyond the Exabyte (Figure 38)[522].

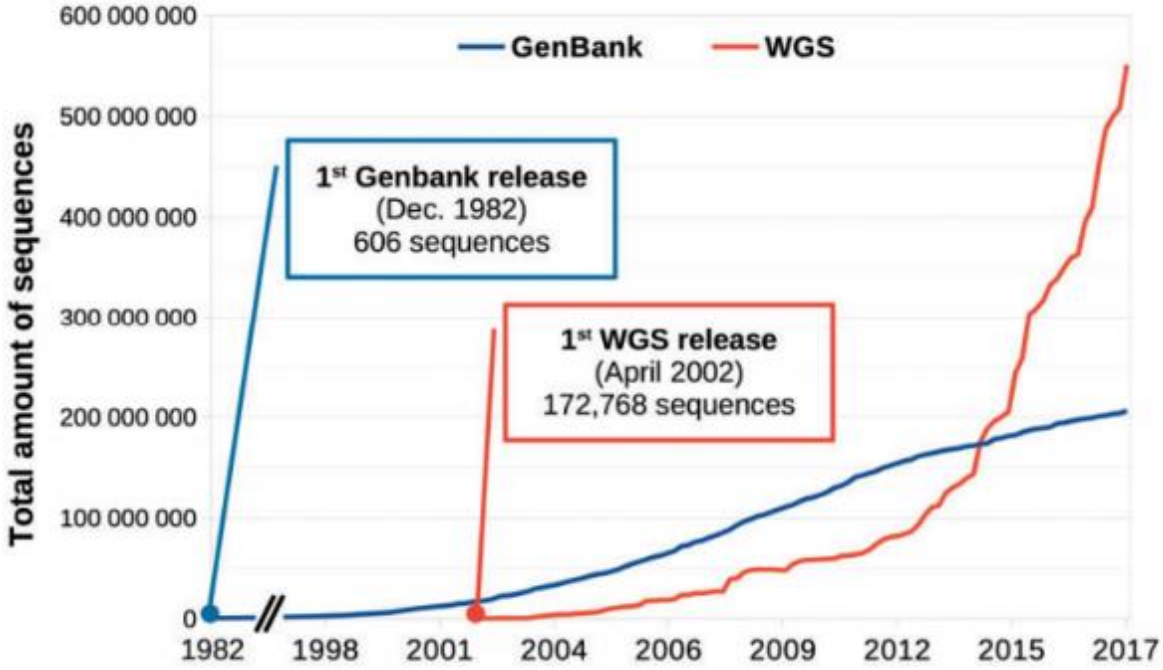


Figure 38: Number of sequences in Genbank and WGS databases over time, from 1982 to the present[504]

The evolution of the sequencing technologies has led to the reconstruction of the genomes of several model organisms and the generation of dedicated databases (e.g., Drosophila[523], Saccharomyces[524], and Human[525]), representing critical resources for the scientific community working on these organisms. These databases contain well-annotated sequences and metadata[504]. More recently, comprehensive genomic databases such as the Sequence Read Archive[526] and the European Nucleotide Archive[527] appeared. These databases store raw sequencing data from different studies, making them available for other reserchers[504].

The development of more user-friendly tools for the data analysis, not requiring a particular background in informatics and programming, like Galaxy, is becoming of paramount importance[528]. Moreover, websites such as SEQanswers[529] and BioStar[530] represent an essential resource, helping scientists to share their work and find help in the community[504].

In the beginning, the sequencing of a single gene required a great deal of hard work and the development of scripts and tools to analyze the data. Today there are many user-friendly tools, allowing the reconstruction of whole genomes using just a desktop computer and short computation time. In conclusion, bioinformatics has become an integral part of life science[504].

5.4.1 Quality control

The output file of high-throughput sequencing methods utilizes the FASTQ format. A typical FASTQ file includes the nucleic acid sequence and the Phred quality score of the base call, both encoded with ASCII charactes[531].

Phred quality score, corresponding to the probability that a base has been erroneously incorporated, was initially developed for the Human Genome Project, to evaluate the quality of the base calls from Sanger sequencing[531]. The software scans the peaks of the chromatogram and scores based on the certainty or accuracy of the call. The scores are logarithmically based, and scores higher than 20 represent greater than 99% accuracy of the base call (Figure 39)[532].

Phred score allows the identification of low-quality reads, during the quality control stage.

One of today's most commonly used tools to perform the quality control of the sequencing data is FASTQC[533], which generates a user-friendly HTML report.

The Oxford Nanopore Technology's (ONT) sequencers generate a sequencing output in FAST5 format. It contains the raw electrical signal levels measured by the nanopores. The FAST5 file structure is based on a typical HDF5 file. The primary data in these files are the "squiggles" that represent pico-amp measurements taken around thousands of times a second at the nanopores. Each read resulting from sequencing a molecule is stored as a single FAST5 file[534].

Poretools toolkit (https://poretools.readthedocs.io/en/latest/) can extract data from FAST5 files (ONT sequencing raw data) and generate FASTQ files form the ONT sequencing raw data[534].

Guppy toolkit can perform both the base-calling and de-multiplexing steps, generating FASTQ files representing different multiplexed samples[535,536].

By using different tools, the FASTQ file is filtered to remove sequencing adaptors or low-quality sequences. This last step can be performed using Cutadapt, which aligns the reads with all adapter sequences, depending on the sequencing platform[537]. Trim Galore[538] (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) is a wrapper tool, including the Cutadapt tool and the QC reports from FASTQC, and thus represents a useful tool for the QC step.

Porechop (https://github.com/rrwick/Porechop), Nanofilt (https://github.com/wdecoster/nanofilt), and Filtlong (https://github.com/rrwick/Filtlong) are specifically designed to perform quality filtering on ONT sequencing data.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Figure 39: Phred quality scores and relatives probabilities of incorrect base call and base call accuracy values[531]

5.4.2 The assembly of sequencing data

Short reads often need to be assembled into more informative sequences. After the removal of adapter sequences, a reference genome can guide the assembly of these sequences. When a reference genome is not present, a more computationally intensive process of *de novo* assembly must take place. With *de novo* assembly, the long reads generated by ONT can help in creating scaffolds for the assembly into contiguous sequences or contigs.

In metagenomics, two different strategies can be applied, based on the aim of the study: read-based and assembly-based approaches. With a read-based approach, a sequence can be aligned to known reference genomes or genes to evaluate coverage and variation. However, this method is not suitable for the discovery of novel organisms. On the other hand, with an assembly-based approach, reads are first *de novo* assembled into contigs and then clustered into “genome bins”, generating scaffolds. Contigs are continuous stretches of sequence without gaps. Scaffolds are created by chaining contigs together, taking into account the relative position and orientation of the contigs in the genome, and may contain gaps. In a metagenomics sample, there is a mix of different genomes from different taxa, and by assembling the sequences, it is possible to reconstruct entire genomes. However, the biological complexity of the sample can make this reconstruction process quite challenging.

The presence of homologous or paralogous sequences can lead to intragenomic or intergenomic chimeric assemblies[539,540]. Different tools are available to assemble reads into large contigs, and most of them are based on *de novo* assembly algorithms[541,542]. The performance of the genome assembly can be evaluated based on the size of the smallest contigs in a set of contigs that makes up at least 50% of the assembly, named N50[541].

The greedy algorithm, which aims for a local optimum, and the graph-based algorithm, which aims for a global optimum, are the two main types of *de novo* assembling algorithms. Graph-based are nowadays the most commonly used type of *de novo* assembly algorithms and are mainly based

on de Bruijn graph assembly. In graph-based algorithms, the sequences are divided into segments of size k, which form a network of overlapping paths leading to the generation of the contigs[542].

In graph theory, k-mers are the nodes, and the overlapping parts are the edges, which can be weighted based on the overlapping length. One of the most recent assembly tools for metagenomics is MetaSPADES[543], which is built upon the commonly used SPAdes genome assembler[478] and combines graphs of different k values. Spades also includes an option to look for circular genomes, a highly important feature in the context of PV detection.

For the ONT sequencing data, after the reconstruction of a sequence, using tools like Nanopolish[500], an improved consensus sequence can be generated. This tool will align the reconstructed draft sequence against the raw sequencing data and will correct potential errors that occurred during the assembling steps, at the signal-level. This step can significantly improve the draft sequence, also considering the high error rate, typical of the raw data from ONT sequencing.

Nanopolish uses a probabilistic model to evaluate whether modifications to the draft sequence (substitutions, insertions, deletions, and substrings) can improve the probability of observing the electric signal data for the collection of MinION reads[500].

5.4.3 Taxonomic classification

In metagenomics studies, the taxonomic classification of the sequences allows the identification of the microbial/viral community profile. There are two main strategies used to perform the taxonomic classification: reference-based and reference-free classifications.

The reference-based classification relies mainly on local alignment tools against a reference database. In contrast, the reference-free classification, which is less common, relies primarily on sequence composition, such as k-mer frequency.

MEGAN is a tool developed in 2016 that can identify sequences obtained after an assembling step (i.e., contigs) or directly short raw reads, not assembled. This tool uses a Blast-based algorithm to align the sequences against a reference database of known sequences[544]. MEGAN assigns the Last Common Ancestor (LCA) to each sequence, using the taxonomic information from the NCBI database.

KRAKEN is a k-mer-based tool for the classification of short sequences, usually obtained through metagenomic studies, allowing a fast classification of the sequences. In this tool, sequences are classified by querying the database for each k-mer in a sequence and then using the resulting set of LCA taxa to determine an appropriate label for the sequence. Sequences that have no k-mers in the database are left unclassified by KRAKEN[545]. A disadvantage of the k-mer- based classifiers is a large amount of memory required.

Centrifuge[546], developed in 2016, represents an improvement of these classification approaches. This tool for the classification of short sequences can analyze data generated in metagenomics studies. In this tool, an indexing scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index, optimized specifically for the metagenomic classification problem, was developed to reduce the memory requirements and the processing time[546].

Moreover, this tool can compute the abundance of each taxon at any taxonomic rank, using a maximum likelihood estimation method.

Some tools have been developed specifically for the taxonomic classification of viral sequences in high-throughput sequencing data.

VirusSeq[547] tool allows the removal of host-genome sequences and then performs the classification of viral sequences present in the sample. VERSE[548] tool can assess the virus genome integration into the host DNA.

Many of these tools, for the taxonomic classification of viral sequences, after a filtering step, perform an identification of the sequences using alignment algorithms like Blastn, against a reference database.

VirusSeeker[549] is a new tool for virus detection and is a BLAST-based NGS data analysis pipeline designed for virome composition description and novel virus discovery. This tool takes advantage of the Blastn algorithm and the NCBI Taxonomy database to identify the viral sequences[549].

Nowadays, there are no specific tools for the identification and classification of PV sequences from high-throughput sequencing data.

Aim of the study

New HPV types are continuously discovered and fully characterized[550,551]. The advent of more and more effective molecular biology technologies make this discovery increasingly rapid.

HPVs are classified based on the L1 nucleotide sequence identity and divided into genera.

The major HPV genera are alpha, beta, gamma, mu, and nu[552]. Many types of the alpha genus have been extensively studied because of their clear involvement in human cancers[200]. This subgroup of viruses, named high-risk HPVs, includes at least 12 HPV types (i.e., HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 and 59) which are the etiological agents of anogenital cancers and a subset of head and neck cancers, particularly oropharyngeal cancer[553,554]. The alpha genus also includes the low-risk types(e.g., HPV6 and 11), associated with benign lesions[555,556].

Although the literature is rich in studies on the characterization of alpha types to understand their role in human diseases, less is known about other HPV genera.

The genus beta includes viruses widely present in the skin of healthy individuals. The majorities of the beta HPV types belong to beta-1 and beta-2 species, but there are other species much less represented. Recent studies showed the potential role of beta HPVs, in combination with UV light exposure, in the development of cSCC[557–567].

The presence of beta HPV types was recently reported in additional anatomical sites other than skin, such as oral mucosal epithelium, eyebrow hairs, penile, external genital samples and, anal canal[471,472,551,568,569].

In particular, species beta-3 HPV types appear to have a dual tropism, being present in the skin and mucosal epithelia[570,571], and biological similarities between beta-3 HPV and mucosal HR HPV types have been identified[567,572].

Gamma genus is represented by a significant number of viruses that have not yet been linked to any malignancy[573]. Several other HPVs from other genera have been discovered, and most of these viruses need to be fully characterized to understand their potential role in human diseases[55,477].

In the past, the major strategies used to discover new HPV types were based on the use of hybridization methods, consensus or degenerate primers, cloning, and Sanger sequencing[382,383,574–580].

Often only a partial sequence, representative of a novel HPV, is reconstructed, and new strategies are required to extend these sequences and obtain the full genome of new viruses.

Therefore, the general aim of this work is the development of high-throughput sequencing-based methodologies for the identification of novel HPV types.

Specific aim #1

The first specific aim of this study was the development of a protocol for the amplification and identification of putative novel HPVs. As a first step, aside from well-validated primer sets, new primers were also designed, based on the HPV L1 genomic region of the known beta types, increasing our chances to expand this HPV genus. After, the different PCR protocols were used to amplify HPV sequences from two kinds of human specimens: skin swabs and oral gargles. Finally, the PCR products were pooled and analyzed using an amplicon targeted sequencing NGS strategy[581].

A pipeline was developed to analyze the NGS data for the identification of papillomavirus sequences.

Specific aim #2

After the identification of putative new HPV types, based on the NGS data analysis, the samples were re-screened to identify the ones positive for specific putative new HPV types. The second specific aim of this work was the reconstruction of the whole genome of one putative new HPV type identified in the previous NGS analysis, for complete genomic characterization.

Specific aim #3

The strategies applied so far for the characterization of HPV genomes are based on the use of outward-directed primers to amplify the whole genome of the virus, then cloned into a vector and sequenced by using a primer-walking Sanger-based strategy[582]. In specific aim #2 of the present

work, we used this approach to characterize new HPV types, but often amplification, cloning, and sequencing steps were laborious and time-consuming.

The third and last specific aim of this study was to determine whether the MinION sequencer can be applied in the genomic characterization of papillomaviruses, with a minimal error rate. At this purpose, the whole genome of the new HPV-ICB2, characterized in specific aim #2, was sequenced using MinION technology. This approach also required the development of a bioinformatics analysis for the reconstruction of the viral genome, using ONT sequencing data.

Materials and methods

Specific aim #1

1.1 Samples collection and DNA extraction

Two different kinds of human specimens were used in the present study, skin swabs and oral rinses. Both skin and oral samples come from previous studies assessing the prevalence of viral DNA[568,583–586].

The VIRUSCAN is a five-year prospective cohort study put in place in 2014 by the Moffitt Cancer Center and the University of South Florida (R01CA177586-01; “Prospective study of cutaneous viral infections and non-melanoma skin cancer”). A 0.9% saline solution was applied to a small area (5 x 5 cm) of the forearm skin, normally exposed to sunlight. Thus a Dacron cotton-tipped swab (Digene, Gaithersburg, MD, USA) was rubbed on the skin a few times to collect exfoliated skin cells. The swab was then placed in a vial with Digene Standard Transport Medium (STM).
A total of 119 skin swabs from the VIRUSCAN study were selected for the present work.

The HPV Infection in Men (HIM) study is a large, multi-national prospective cohort study of the natural history of HPV infection in men[587–590]. A total of 62 oral rinses from the HIM study were selected for the present work.

Additional 85 oral samples were selected from a pilot study on the prevalence of Helicobacter Pylori infection in the Latvian population. This study was approved (No. 8-A/15) by the Ethics Committee of Riga East University Hospital Support Foundation.
The DNA extraction was performed using the EZ1 Advanced XL machine (Qiagen) and the EZ1 DNA Tissue Kit. After, all the samples were analyzed at the International Agency for Research on Cancer (Lyon, France) for HPV-DNA positivity.

1.2 PCR protocols

Six different PCR protocols were used in this study, all amplifying a portion of the L1 region of the HPV genome.

The Beta3-1 is a mix of 11 consensus beta-3 primers, while the Beta3-2 is a mix of 4 broad spectrum beta-3 degenerated primers. Both these primers mix amplify a fragment of roughly 450bp (Figure 40) and have been developed in our laboratory. These primers were synthesized by MWG Biotech (Ebersberg, Germany) and mixed to obtain a 10X solution containing 2µM of each primer. PCRs were performed with the QIAGEN Multiplex PCR kit (Hilden, Germany) according to the manufacturer's instructions.

CUT primers are a mix of 5 broad spectrum cutaneous degenerate primers previously described, amplifying a portion of the L1 region of around 370bp (Figure 40 and Table 5)[591].

FA-types (FAP) primers are a mix of 2 primers, one forward and one reverse, amplifying a portion of the L1 region of roughly 478bp (Figure 40 and Table 5)[592].

Also, a new set of FAP primers was developed in our laboratory, i.e., FAP59.1, FAP59.2, and FAP64.1 (Table 5), amplifying the same region of the L1 targeted by the original FAP primers (Figure 40). FAPM1 is a mix of 5 primers, including both the original and the novel FAP primers (Table 5). FAPM2, rather, is a mix of just 2 of the new FAP primers (Table 5). The PCR conditions used for FAPM1 and FAPM2 were the same as for the original FAP protocol.

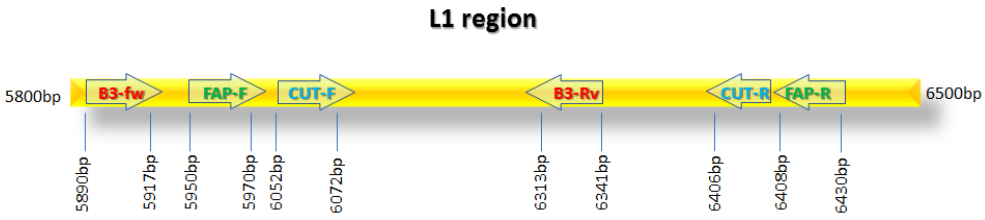


Figure 40: Each of the primers sets used in this study amplifies a portion of the L1 gene of the HPV genome: Beta3-1 and Beta3-2 [~450 bp]; FAP* [478 bp]; CUT [370 bp]

Primer mix	Primer sequence (5'-3')
Beta3-1	
B3L1FW3	AGGACATCCATACTTTGAGGTTCGAG
B3L1FW4	TAGGACATCCATATTTTGATGTGAGAG
B3L1FW5	GATGTTAGAGACACTGGAGATTCAACA
B3L1FW6	GATGTTAGAGACACTGGGGATTCAACA
B3L1FW7	GATGTTAGAGACACTGTGGATCAAACA
B3L1RW	ATAATAGTATTTCTTAATTCTAATGGAGG
B3L1RW4	ATAACTGAATTGATTAATTCTAATGGAGG
B3L1RW5	ATAACTGTATTTACTAATTCTAAAGGTGG
B3L1RW6	TACAGTATTTACCAGTTCCAAAGGTGG
B3L1RW7	ATTACAGTATTAACTAATTCTAAAGGTGG
B3L1RW8	ATTACAGTATTTACTAATTCTAAAGGTGG
Beta3-2	
B3L1FW1	GTAGGACATCCATAYTTTGAKGTKiGAG
B3L1FW2	TTGATGTTAGAGACACTGiDGATYMAACA
B3L1RW1	ATAAiWGWATTKYTTAATTCTAATGGAGG
B3L1RW2	ATTACAGTATTiACKARTTCYAAAGGTGG
CUT	
CUT1Fw	TRCCiGAYCCiAATAARTTTG
CUT1AFw	TRCCiGAYCCiAACAGRTTTG
CUT1BFw	TRCCiGAYCCiAAtAGRTTTG
CUT1CFw	TRCCiGAYCCiAACAARTTTG
CUT1BRv	ARGAYGGiGAYATGGTiGA
FAP	
FAP59	TAACWGTIGGICAYCCWTATT
FAP64	CCWATATCWVHCATITCICCATC
FAPM1	
FAP59.1	TAACAGTDGGiCAYCCWTWTT
FAP59.2	TAACAGTDGGiCAYCCWTAYT
FAP64.1	CCDATATCWVHCATATCiCCATC
FAP59	TAACWGTIGGICAYCCWTATT
FAP64	CCWATATCWVHCATITCICCATC
FAPM2	
FAP59.2	TAACAGTDGGiCAYCCWTAYT
FAP64.1	CCDATATCWVHCATATCiCCATC

Table 5: *Oligonucleotide sequences* and composition of *primers mixes used in this study* (i = Inosine; W= A or T; D = A, G or T; K = T or G; Y = C or T; M = A or C).

1.3 Validation of the new set of primers

The new PCR primers, developed in our laboratory, were tested to evaluate their specificity and sensibility. For this purpose, an artificial mixture composed of a known cloned HPV and human genomic DNA was used at different relative concentrations of the viral genome (10-fold dilution series starting from 10.000 to 0 copies of the cloned virus). The PCR products were loaded on a 2% agarose gel and analyzed after gel electrophoresis.

1.4 NGS analysis

The QIAquick gel extraction kit (Qiagen®, Hilden, Germany) was used to purify the PCR products according to the manufacturer's instructions. After, the Agencourt AMPure XP PCR purification kit was used with a 1.8x beads ratio, to remove residual impurities and contaminants from previous steps.

Finally, eight pools of around 50 samples each were generated, based on the PCR protocol and the kind of human specimen used (Table 6).

Nextera XT DNA Library Preparation Kit (Illumina Inc., San Diego, CA, USA) was used for the library preparation while, for the indexing of the pools Illumina MiSeq, double indexed adapters (Illumina Inc., San Diego, CA, USA) were selected.

AMPure XP beads were used again to clean-up the indexed libraries after the amplification and ligation steps. In this case, a 1x beads ratio was used to increase the average size of the libraries, removing short fragments.

After the purification step, the quality and average size of the newly generated libraries were assessed by using chip-based capillary gel electrophoresis (Agilent 2100 Bioanalyzer).

Library generation and clean-up steps were repeated several times per each of the PCR pools to optimize the protocol, allowing the selection of libraries of around 200-300 bp in average size, then used as input DNA for the sequencing step.

Next-generation sequencing was performed on Illumina Miseq machine, using a sequencing kit v3 (600 cycles) and 10% of PhiX (Illumina, San Diego, California, USA) that helps to sequence low-diversity samples like our PCR products.

PCR pools	PCR protocols	Specimens	N	NGS pools
1	Beta-3-1	Skin swab	41	1
2	Beta-3-2		9	
3	FAP		52	2
4	FAPM1		54	3
5	CUT		57	4
6	FAPM2	Oral gargle	43	5
7	FAPM1		56	6
8	CUT		55	7
9	Beta 3-1		9	8
10	Beta 3-2		4	
11	FAP		11	
12	FAPM1		11	
13	FAPM2		12	

Table 6: Organization of the PCR and next-generation sequencing pools, based on the primers used and the kind of specimen.

1.5 Bioinformatics analysis

FASTQC (v0.11.5)[533] and MultiQC (v1.0)[593] were used to perform quality control of the raw sequencing data. Trim Galore (v0.4.4)[538] was used to remove low-quality reads-ends and remaining adapter sequences. VSEARCH (v2.4.0)[594] was used to merge forward and reverse reads, de-replication, *de novo* identification of chimeric sequences, and sequences clustering. Thus, the clusters of sequences generated were identified in a local server by aligning them to the NCBI Nucleotide collection (nr/nt) database (updated on March 2017) using Megablast in the Blast package (v2.6.0+)[480].

Based on the MegaBlast results, the reads were clustered (based on the E-value), and then each cluster was processed using CAP3 assembling program[595].

A total of 458 full-L1 ORF nucleotide HPV sequences from the PaVe database, updated in January 2018 (https://pave.niaid.nih.gov/)[171] were used to generate a reference species phylogenetic tree.

The L1 nucleotide sequences were aligned using the MUSCLE algorithm, with default parameters[596], in MEGA7[597]. The final full-length L1-ORF alignment encompassed 458 full L1-ORF nucleotide sequences, 2259 positions, and 627 distinct alignment patterns.

The best substitution model and the phylogenetic inference were retrieved using MEGA7. The codon positions included were 1st + 2nd + 3rd + noncoding, and based on the alignment, the partial deletion parameter (all positions with < 95% site coverage discarded) was selected to enable the inclusion of taxa with potential missing data. The final dataset included 1383 positions. Five rate categories were used for the discrete gamma distribution (parameter = 1.0326). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 2.5307% sites). Initial trees for the heuristic search were obtained automatically by applying neighbor-joining, and BioNJ algorithms to a matrix of pairwise distances estimated using the maximum composite

likelihood (MCL) approach and then selecting the topology with the superior log-likelihood value (−389774.5274).

A generalized time-reversible (GTR) model was used in MEGA7 with 500 bootstrap replicates for the phylogenetic inference[598].

The consensus sequences generated by CAP3 were mapped using PaPaRa[599] against the fixed reference multiple sequence alignment (MSA).

Then, the sequences were mapped into the reference species phylogenetic tree using the evolutionary placement algorithm (EPA) in RAxML (v8.2.11)[600,601] using the same nucleotide substitution model used to infer the phylogenetic reference tree.

A script was developed in our laboratory to parse the output format[602] of the EPA.

For better identification of the sequences, in parallel, the contigs were aligned against the whole PaVe database, including 330 PVs genomes at the time of the analysis (updated on March 2017), using the Blastn algorithm in the Blast package (v2.6.0+). Krona tool[603] was used for the graphical representation of the data.

2 PVampliconfinder

PVAmpliconFinder uses alignment similarity metrics, in combination with molecular evolution time for improved identification and taxonomic classification of novel PVs.

PVAmpliconFinder takes paired-end FASTQ files as input and, as the first step, using TrimGalore[538], performs a quality trimming of the raw sequencing data, removing adapter and primers sequences (Figure 41 A). This step also discards low-quality bases, sequences of less than 32 bp, poly-A sequences, and reads with low average quality score.

The quality of the sequences is assessed before and after the trimming step, using FASTQC[533] and MultiQC[593].

Then, merging of the paired reads, de-replication, de novo chimera detection, and *de novo* clustering steps are performed using VSEARCH[594] tool. The merging step is followed by the generation of a FASTQC report for a quality check, enabling the identification of primer contamination (Figure 41 B).

The de-replication step collapses identical sequences into a single template keeping the information on the number of reads used to form the final template, reducing the data complexity.

The *de novo* chimera detection step identifies and removes chimeric DNA sequences that can occur during PCR amplification, especially when sequencing a unique region.

The *de novo* clustering step groups the sequences sharing more than 98% level of identity (default value), and the resulting unique sequence is used for the downstream analysis.

Follows the identification of the PV-related sequences (Figure 41 C) trough alignment against the complete NCBI “nt” nucleotide sequence database, which includes all sequences from all species (Figure 41 C). Next, groups of sequences are defined based on two characteristics: the best MegaBlast subject sequence for each query, and the percentage of similarity of each sequence with its corresponding best subject sequence. As a result of this process, the sequences are divided into two groups: putative known PVs and putative new PVs (Figure 41 C).

The grouped sequences are then *de novo* assembled using CAP3[595] *de novo* assembling tool, extending the sequence lengths to cover the full L1 region targeted by the different primer systems used in the PCR step (Figure 41 D).

The extended PV sequences are classified using two methods. The first is based on the taxonomic classification of the best subject match (using the e-value computed by Blastn[604]) when aligned against a comprehensive database of PV sequences (the full L1 gene nucleotide sequence database available in the PaVE[171] database). The second is based on molecular

evolution using the Randomized Axelerated Maximum Likelihood-Evolutionary Placement Algorithm[601] (RaxML-EPA) (Figure 41 E).

For the RaxML-EPA-based taxonomical classification, a phylogenetic reference tree is constructed based on the full-L1 ORF nucleotide sequences of the available PV genomes present in the PaVE database.

Phylogenetic inference is performed using MEGA7[597] tool. The Parsimony-based Phylogeny-Aware Read alignment (PaPaRa)[605] program is used to align each contig sequence, reconstructed during the previous de novo assembly step (Figure 41 D), against the MSA[599].

Subsequently, the evolutionary placement algorithm (EPA) (20) in RaxML[601] is run to place the sequences into the reference tree (Figure 41 E), based on PaPaRa[605] multiple alignments. The EPA is run using the same nucleotide substitution model used to infer the phylogenetic reference tree.

Several output reports are generated as Excel files, FASTA files, or graphical images from the different steps of the workflow.

A list of fully characterized putative new Papillomaviridae sequences and graphical representations of the relative abundance and diversity of HPV sequence detected in the tested samples are generated.

The PVAmpliconFinder workflow and its source code are freely available on the GitHub platform: https://github.com/IARCbioinfo/PVAmpliconFinder.

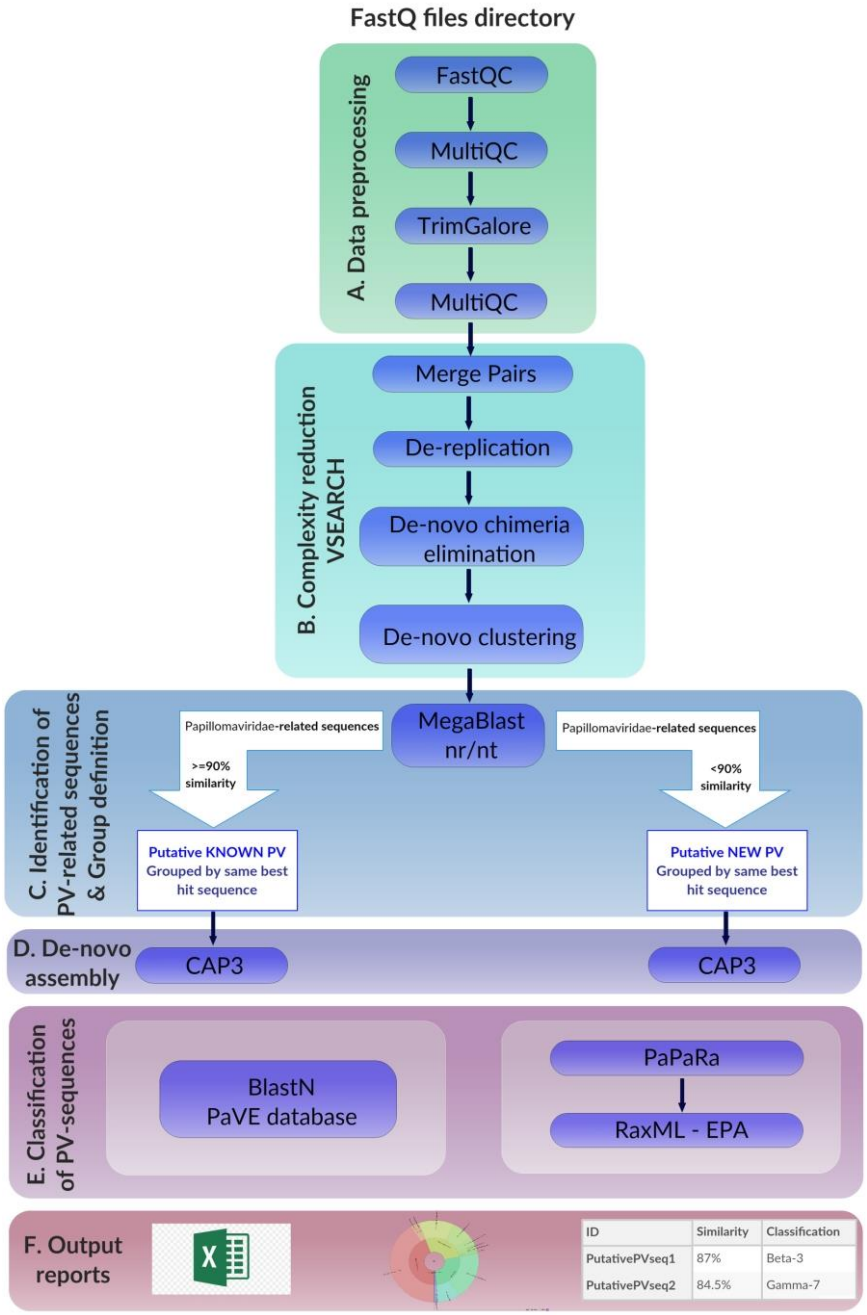


Figure 41: PVAmpliconFinder workflow

Specific aim #2

1. Rolling circle amplification and search of the novel HPV types

According to the NGS data analysis, the sequences representative of putative new HPV types were selected and used to design primers for the screening of the original samples to detect the ones positive for the specific putative novel HPV types. The primers for the detection of the novel HPV types were designed based on little portions of the L1 gene that were retrieved from the NGS data analysis.

Primers were synthesized by MWG Biotech (Ebersberg, Germany) and used with a concentration of 0.2 µM in the PCR reaction. PCRs were performed with the QIAGEN Multiplex PCR kit (Hilden, Germany) according to the manufacturer's instructions.

Furthermore, before the screening, to enrich the samples of viral genome copies, rolling circle amplification (RCA) was performed on the original DNA specimens using Illustra TempliPhi 100 RCA kit (GE Healthcare Life Sciences, Little Chalfont, Buckinghamshire, UK) as described by Schowalter *et al*. .[606] with supplementation of 450 µM dNTPs as described by Rector *et al.*[607].

2. Long-distance PCR, cloning and Sanger sequencing

After the identification of the sample positive for a specific putative novel HPV type, outward-directed PCR was performed for the amplification of the whole genome of the virus. The primers were designed based on the L1 sequences obtained from the NGS.

TaKaRa LA Taq® Hot Start (TAKARA Bio Inc., Kusatsu, Japan) enzyme was used to perform long-distance PCR, in agreement with the manufacturer's instructions. After, the PCR products were

cloned using Topo XL PCR cloning kit® (Thermo Fisher Scientific, Waltham, USA), in agreement with the manufacturer's instructions.

Especially for some portion of the viral genome, where the sequence identity was uncertain since potential errors could be introduced by the reaction with the TaKaRa enzyme, additional PCRs were performed using AmpliTaq Gold™ DNA proofreading enzyme, (Thermo Fisher Scientific, Waltham, USA).

Sanger sequencing analyses were conducted by the GATC Biotech company (Costanza, Germany). This sequencing service uses cycle sequencing technology (dideoxy chain termination/cycle sequencing) on an ABI 3730XL sequencing machine. The assembling of the sequences was carried out using CAP3 DNA sequence assembly program[595].

Specific aim #3

1. Human specimen

In the previous steps of the present work, the strategies used for the identification of novel HPV types allowed the discovery of HPV-ICB2 (accession number MK080568), a novel beta-2 HPV type with a genome of 7441 bp in length.

This virus has been isolated from a human forearm skin swab sample and fully characterized in our laboratory. The sample was originally collected for the VIRUSCAN study mentioned above.

2. DNA extraction and Rolling circle amplification

The DNA extraction of the swab specimen was conducted using QIAquick PCR purification columns (Qiagen). The purified DNA was digested with *NotI* (NEB) and Plasmid Safe (exonuclease V, Epicentre), then ethanol precipitated. The pelleted DNA was re-suspended and amplified using Illustra TempliPhi 100 RCA kit (GE Healthcare Life Sciences, Little Chalfont, Buckinghamshire, UK) as described by Schowalter *et al.* .[606] with supplementation of 450 µM dNTPs as described by Rector *et al*.[607].

3. MinION library preparation

The SQK-LSK109 protocol for 1D PCR barcoding amplicons was followed to generate three independent libraries (Figure 42).

The entire genome of HPV-ICB2 was amplified form a 1/100 dilution of the RCA product, using KAPA HiFi HotStart ReadyMixPCR kit following the manufacturer's instructions (KAPA

biosystems, Boston, MA, USA). HPV-ICB2-specific outward-directed primers tailed with MinION adapters were used for the amplification.

The sequences of the tailed primers used are: forward primer, 5'-**TTT CTG TTG GTG CTG ATA TTG C**CA GAC AGA ACA CAT CTT TTG ATC C-3' and reverse primer, 5'-**ACT TGC CTG TCG CTC TAT CTT C**TC GTC CCG TGA CCC ACC CTG A-3'. These primers were used at a final concentration of 1.6 pM each.

These primers were designed based on the NGS sequencing data obtained in aim #1 of the present work and specifically on an L1 fragment of 99 bp.

Only proofreading polymerases were used to avoid the introduction of errors during the amplification steps. A C1000 Touch thermal cycler (Bio-Rad Laboratories, Inc.) was used for the amplification, with the following protocol: an initial denaturation step of 3 minutes at 95°C followed by 35 cycles of denaturation at 98°C (20 seconds), annealing at 64°C (15 seconds), and extension at 72°C (8.5 minutes), with a final extension at 72°C (10 minutes) to generate a 7485-bp product. Then, purification of the PCR product was conducted using a QIAquick gel extraction kit (Qiagen, Hilden, Germany) following the manufacturer's instructions.

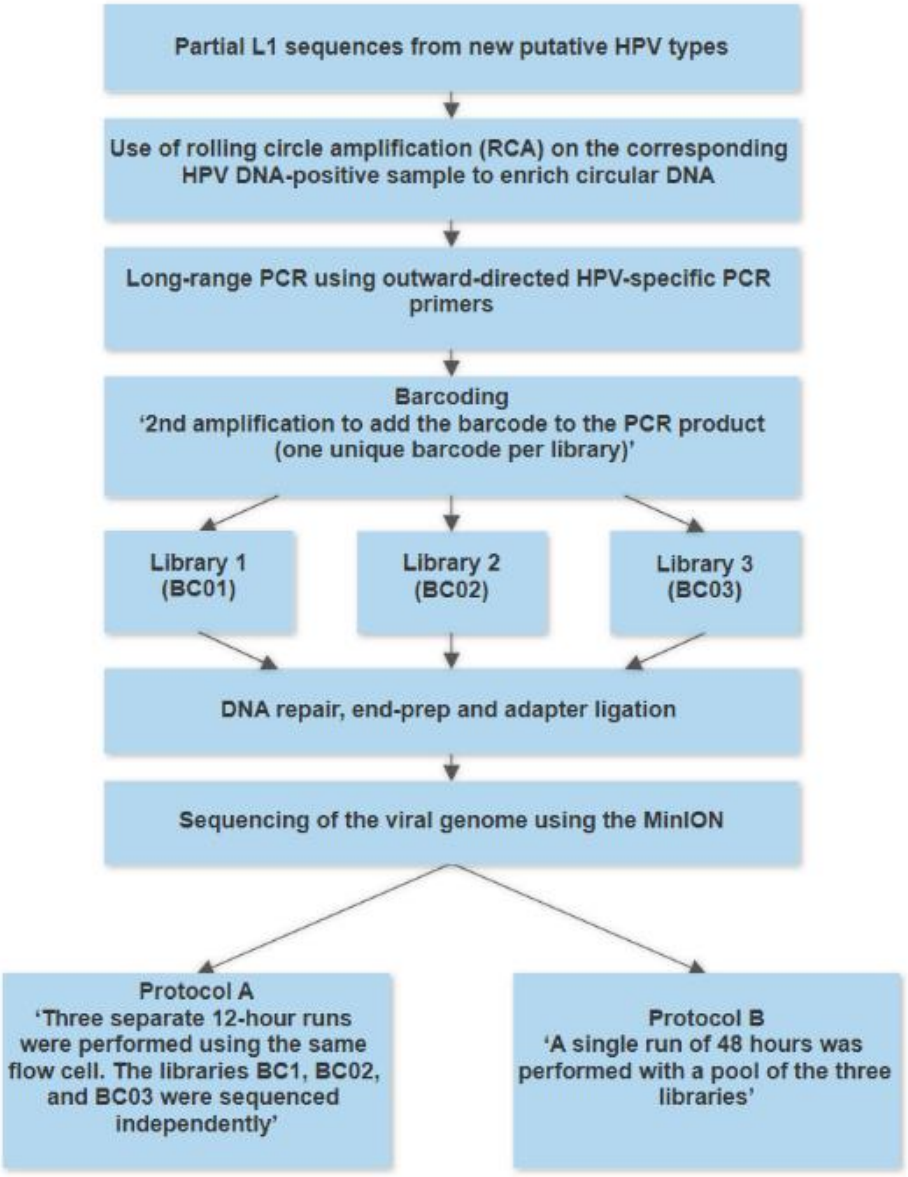


Figure 42: Workflow strategy applied to sequence the HPV genome using MinION technology

4. Barcoding

A second PCR was performed, adding sequencing barcodes to generate three different libraries from the three PCR products.

KAPA HiFi HotStart ReadyMixPCR kit (KAPA Biosystems, Boston, MA, USA) was used for this purpose.

The barcodes used are BC01, BC02, and BC03 from the PCR Barcoding Expansion 1-12 kit (EXP-PBC001, Oxford Nanopore Technologies, Oxford, UK).

The PCR was performed with the following conditions: an initial denaturation step of 3 minutes at 95°C was followed by 35 cycles of denaturation at 98°C (20 seconds), annealing at 62°C (15 seconds), and extension at 72°C (8,5 minutes), with a final extension at 72°C for 8,5 minutes.

The PCR was followed by a purification step using Agencourt AMPure XP beads (Beckman Coulter, Pasadena, California, USA) with a 1.8x beads to sample volume ratio following the manufacturer's instructions. Then, the purified product was resuspended in 40 µl of Invitrogen™ UltraPure™ DNase/RNase-Free distilled water (Gibco, Life Technologies, Paisley, UK).

5. MinION sequencing run

Following the SQK-LSK109 protocol (Oxford Nanopore Technologies, Oxford, UK), DNA repair, end-prep, adapter ligation, and clean-up steps were performed. The adapted and purified DNA was then quantified using Qubit fluorometer (Thermo Fisher Scientific, Waltham, USA) and loaded into the MinION sequencing flow cell according to the manufacturer's instructions (Oxford Nanopore Technologies, Oxford, UK). The kind of flow cell used for this experiment is the FLO-MIN106D. As shown in figure 42, two different sequencing protocols were used. In protocol A, three consecutive runs of 12 hours each were performed sequencing the three libraries separately (BC01, BC02, and BC03). Finally, a run of 48 hours was performed by pooling all the three libraries together.

After each run, the flow cell was cleaned using a Flowcell Wash Kit (Oxford Nanopore Technologies Oxford, UK) following the manufacturer's instructions.

6. Bioinformatics data analysis

SAMtools was used to calculate the coverage and Pysamstats v1.1.2 for the percentage of similarity by considering the number of matches and mismatches per each base (https://github.com/alimanfoo/pysamstats)[608].

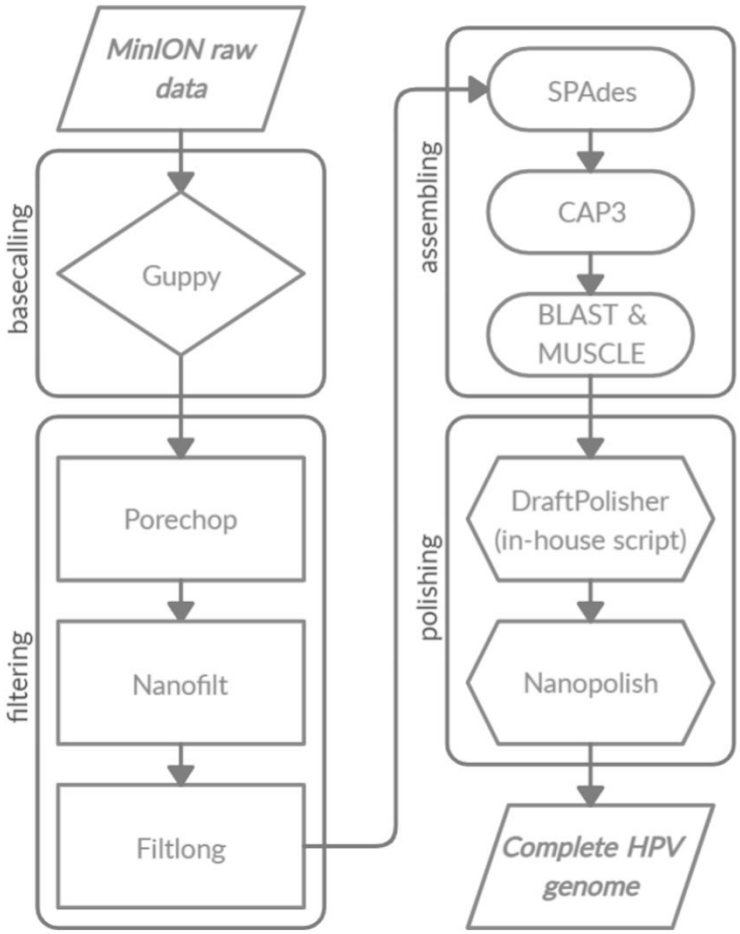


Figure 43: Bioinformatics pipeline used in this work for the analysis of the MinION sequencing data and thus the reconstruction of a complete HPV genome

Base-calling and de-multiplexing were performed using Guppy v3.1.5+ (https://community.nanoporetech.com/protocols/Guppy-protocol)[609,610] with High Accuracy Calling (HAC) configuration (Figure 43). Then, Porechop v0.2.4 (https://github.com/rrwick/Porechop) was used to remove adapter sequences from the raw reads

and NanoFilt v2.2.0 (https://github.com/wdecoster/nanofilt)[611] to extract long (>200 bp) and high-quality reads. Additional size and quality filtering steps were performed using Filtlong v0.2.0 (https://github.com/rrwick/Filtlong), removing reads shorter than 200 bp and low-quality reads (Figure 43).

After, consensus sequences were generated using SPAdes v3.10.1[612]. Two BLAST analyses were performed to assess the effectiveness of the filtering steps, one using the Megablast algorithm, after the base-calling, and another using Blastn on the contigs generated by SPAdes (Figure 43).

The second BLAST analysis allowed the identification of the closest HPV type that will be used as a reference for the filtering of the sequences.

CAP3 assembling tool[613] was used for a second assembling level, generating longer consensus sequences (Figure 43).

Porechop v0.2.4, NanoFilt v2.2.0, Filtlong v0.2.0, SPAdes, and CAP3 tools were added in a script named “MinION_reads_filtering_pipe_steps2to6.py” (https://github.com/IARCbioinfo/MinION_pipes).

BLASTn from the BLAST v2.9.0+ package[604,614,615] and MUSCLE v3.8.1551[616] were used to guide the assembling of the consensus sequences representing the virus, using the closer HPV type identified earlier, as reference.

At this point, a draft genomic sequence, representative of the whole genome of the virus was obtained. Draftpolisher cov v1.0 with default parameters was then used to remove potential errors that occurred in the assembling steps (Figure 43). Draftpolisher is an in-house script, developed in our laboratory, performing an alignment of the draft sequence with a reference genome, using the MUSCLE alignment tool, to identify the mismatches and gaps between the two sequences. Then, the tool extracts *k*-mers of size 2**k*+1 (*k* default parameter = 8 bp) representative of each

of the mismatches and gaps, from both the query sequence (draft sequence) and the subject sequence (the “reference” genome).

At this point, the tool interrogates a database of sequences represented by the SPAdes assembling output, and the k-mers frequency (i.e., the number of occurrences of each *k*-mer) is evaluated for both draft and reference sequences. Per each not matching position, the k-mer frequency is evaluated comparing the draft and reference sequence, and this is used to generate the final polished sequence.

In the case of equal frequencies, the nucleotide from the draft sequence is selected by default.

Draftpolisher acts at the consensus level and generates a polished sequence in a few minutes.

The second step of signal-level polishing is performed using Nanopolish v0.11.0(https://github.com/jts/nanopolish)[609], (Figure 43).

For this last step of polishing, a Nanopolish-based pipeline was developed in our laboratory, including Minimap2 v2.15 (https://github.com/lh3/minimap2)[617] and SAMtools v1.9 (https://github.com/samtools)[618] together with Nanopolish, for the processing of the sequence.

Nanopolish was used to improve the assembly using the FAST5 raw data as a reference. Minimap2 and SAMtools were used to align the reads to the draft sequence and to sort alignments, respectively.

A script named “nanopolish_pipe_step9.py” including this Nanopolish-based pipeline (https://github.com/IARCbioinfo/MinION_pipes) was developed in our laboratory.

The RCA and PCR steps used to amplify the viral genome produced a large number of redundant sequences, making the Nanopolish step problematic because of the characteristics of the algorithm. To reduce the number of redundant sequences to be processed by Nanopolish, the raw FASTA file of the sequences was divided into 200 smaller files by using Fasta-splitter v0.2.4.

Thus, a loop was introduced to process each of the files generated. This strategy was also used to reduce processing time.

Finally, the “nanopolish_pipe_step9.py” script generates a consensus sequence “final_consensus.fasta” by assembling the 200 polished draft sequences using CAP3.

The output sequence of this process was loaded on the SRA database, together with the MinION raw sequencing data.

Before and after the polishing step, BLAST analysis was performed to evaluate the effectiveness of the process. Finally, the complete genomes generated from each of the first three runs were aligned to generate a final consensus sequence using the MUSCLE alignment tool.

Raw sequencing data are available at the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra), under BioProject PRJNA602805. The workflow and the tool can be found at the IARC bioinformatics platform on GitHub (https://github.com/IARCbioinfo).

Results

Specific aim #1

1. Isolation of novel HPV types using broad-spectrum primers and NGS

Novel and well-validated sets of primers, targeting the HPV L1 ORF were used in combination with NGS for the detection of putative novel HPV types. Two types of human specimens were analyzed, skin swabs and oral gargles.

Published: **Rosario N Brancaccio**, Alexis Robitaille, Sankhadeep Dutta, Cyrille Cueni, Daiga Santare, Girts Skenders, Marcis Leja, Nicole Fischer, Anna R Giuliano, Dana E Rollison, Adam Grundhoff, Massimo Tommasino, Tarik Gheit. Generation of a Novel Next-Generation Sequencing-Based Method for the Isolation of New Human Papillomavirus Types. Virology. 2018 Jul; 520:1-10

1.1 Design and validation of the novel HPV PCR primers

New primers were designed to identify putative new HPV types. We focused our research on the identification of new beta HPV types, and thus known beta types have been selected and aligned to define consensus sequences.

As a first approach, the L1 open reading frame (ORF) of the known beta-3 HPV types (HPV49, 75, 76, 115) was considered and aligned using the Clustal W2 multiple sequence alignment tool[619], allowing the generation of two new sets of primers for the amplification of HPV sequences, (i) a mix of 11 specific primers named beta3-1 and (ii) a second mix of broad-spectrum degenerated primers named beta3-2. The composition of the new primers set is shown in table 5.

The second approach was to modify the original FAP primers[592], developed in 1999, representative of 77 HPV types from different genera[620,621], to generate new broad-spectrum degenerate primers.

MUSCLE (3.8) multiple sequence alignment tool[622] was used to align 46 sequences representative of all the beta HPV types known in March 2017 according to the PaVe database (https://pave.niaid.nih.gov/)[171], (Table 7).

Therefore, three new broad-spectrum degenerate primers, namely FAP59.1, FAP59.2, and FAP64.1, were generated based on the original FAP primers sequences. The original FAP primers were modified according to the alignment of the L1 sequences from all the known HPV beta types (Table 7). These new primers, while keeping a general broad-spectrum nature, allowing the identification of a broad spectrum of HPVs, are slightly more beta-types oriented.

Two different mixtures were generated using these new primers. The FAPM1 mix includes the original FAP primers (i.e., FAP59 and FAP64) and the new ones (i.e., FAP59.1, FAP59.2, and FAP64.1) while the FAPM2 is a mix of FAP59.2 and FAP64.1, two of the new primers generated (Table 5).

To test the specificity and sensibility of the primers, artificial mixtures of known HPV types cloned genomes and human genomic DNA, were used to mimic the experimental conditions.

The Beta-3 protocol allowed the detection of beta-3 HPVs with a limit of detection of 10 copies.

FAPM1 protocol allowed the detection of beta-2 and beta-3 HPV types with a limit of detection of 10 copies.

FAPM2 protocol allowed the detection of beta-2 HPV types with a limit of detection of 10.000 copies, while for beta-3 the detection limit was 10 copies. As an example of the sensitivity and specificity test performed on the new sets of primers, figure 44 shows the results obtained for FAPM1 protocol using a 10-fold dilution series of HPVs 38 and 49.

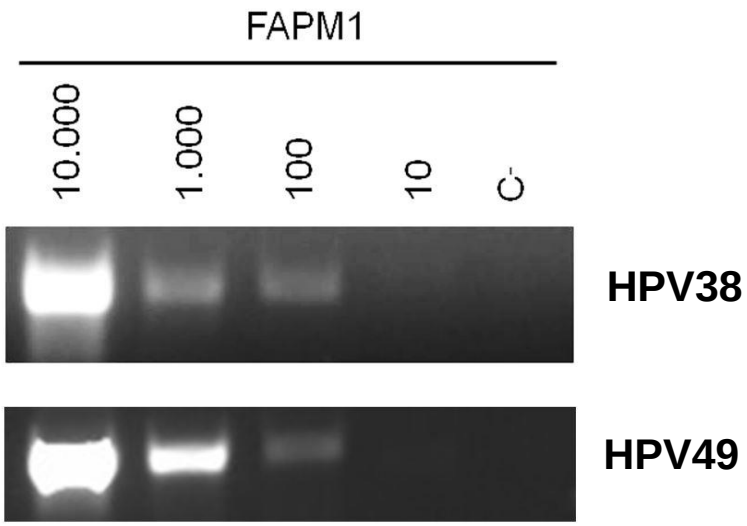


Figure 44: Sensitivity and specificity test for FAPM1 PCR protocol. 10-fold dilutions of HPV38 and 49 were used to evaluate the effectiveness of the FAPM1 PCR protocol. PCR products were analyzed by electrophoresis on a 2% agarose gel

A

	FAP59	T	A	A	C	W	G	T	I	G	G	I	C	A	Y	C	C	W	T	A	T	T
HPV TYPE	A.N.																					
5	M17463	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	M12737	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	X74464	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	X74466	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	X74467	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	X74468	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
17	X74469	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.
19	X74470	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	U31778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	U31779	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	U31780	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
23	U31781	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
24	U31782	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
25	X74471	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
36	U31785	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
37	U31786	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	U31787	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
47	M32305	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
49	X74480	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
75	Y15173	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
76	Y15174	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
80	Y15176	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
92	AF531420	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
93	AY382778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
96	AY382779	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
98	FM955837	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
99	FM955838	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
100	FM955839	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
104	FM955840	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
105	FM955841	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
107	EF422221	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
110	EU410348	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
111	EU410349	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
113	FM955842	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
115	FJ947080	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
118	GQ246951	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
120	GQ845442	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
122	GQ845444	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
124	GQ845446	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
143	HM999995	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
145	HM999997	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
150	FN677755	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
151	FN677756	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.
151	JF304768	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
152	HE963025	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
174	HF930491	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

	FAP59.1	T	A	A	C	A	G	T	D	G	G	i	C	A	Y	C	C	W	T	W	T	T
	FAP59.2	T	A	A	C	A	G	T	D	G	G	i	C	A	Y	C	C	W	T	A	Y	T
HPV TYPE	A.N.																					
5	M17463	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	M12737	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	X74464	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	X74466	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	X74467	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	X74468	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	X74469	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
19	X74470	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	U31778	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	U31779	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	U31780	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
23	U31781	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
24	U31782	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
25	X74471	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
36	U31785	.	G	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
37	U31786	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	U31787	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
47	M32305	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
49	X74480	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
75	Y15173	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
76	Y15174	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
80	Y15176	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
92	AF531420	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
93	AY382778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
96	AY382779	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
98	FM955837	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
99	FM955838	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
100	FM955839	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
104	FM955840	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
105	FM955841	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
107	EF422221	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
110	EU410348	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
111	EU410349	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
113	FM955842	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
115	FJ947080	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
118	GQ246951	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
120	GQ845442	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
122	GQ845444	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
124	GQ845446	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
143	HM999995	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
145	HM999997	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
150	FN677755	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
151	FN677756	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.
152	JF304768	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
159	HE963025	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
174	HF930491	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

B

HPV TYPE	FAP64 A.N.	C G	C A	W T	A G	T G	A I	T G	C A	W I	V A	H T	C G	A D	T B	I W	T G	C A	I T	C A	C T	A W	T G	C G
5	M17463	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
8	M12737	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	X74464	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	X74466	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	X74467	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	X74468	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	X74469	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
19	X74470	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	U31778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	U31779	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	U31780	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
23	U31781	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
24	U31782	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
25	X74471	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
36	U31785	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
37	U31786	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	U31787	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
47	M32305	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
49	X74480	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
75	Y15173	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
76	Y15174	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
80	Y15176	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
92	AF531420	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
93	AY382778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.
96	AY382779	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
98	FM955837	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
99	FM955838	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
100	FM955839	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
104	FM955840	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	C	.	.	.	.	.
105	FM955841	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
107	EF422221	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
110	EU410348	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
111	EU410349	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
113	FM955842	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
115	FJ947080	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
118	GQ246951	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
120	GQ845442	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
122	GQ845444	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
124	GQ845446	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
143	HM999995	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
145	HM999997	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
150	FN677755	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
151	FN677756	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
152	JF304768	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
159	HE963025	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
174	HF930491	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

HPV TYPE	FAP64.1 A.N.	C G	C A	D T	A G	T G	A i	T G	C A	W T	V A	H T	C G	A D	T B	A W	T G	C A	I T	C A	C T	A H	T G	C G
5	M17463	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
8	M12737	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	X74464	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	X74466	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	X74467	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	X74468	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	X74469	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	C	.	.	.	.	.
19	X74470	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	U31778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	U31779	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	U31780	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
23	U31781	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	C	.	.	.	.	.
24	U31782	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
25	X74471	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
36	U31785	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
37	U31786	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	U31787	.	.	C	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
47	M32305	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	C	.	.	.	.	.
49	X74480	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
75	Y15173	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
76	Y15174	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
80	Y15176	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
92	AF531420	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
93	AY382778	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
96	AY382779	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	C	.	.	.	.	.
98	FM955837	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
99	FM955838	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
100	FM955839	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
104	FM955840	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	C	.	.	.	.	.
105	FM955841	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
107	EF422221	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
110	EU410348	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
111	EU410349	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
113	FM955842	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
115	FJ947080	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
118	GQ246951	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
120	GQ845442	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
122	GQ845444	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
124	GQ845446	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
143	HM999995	.	.	C	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
145	HM999997	.	.	C	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
150	FN677755	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
151	FN677756	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
152	JF304768	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
159	HE963025	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
174	HF930491	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 7: Design of the new FAP primers based on the alignment of the original FAP primers with 46 known beta types using MUSCLE alignment tool. (A) Forward primer, (B) Reverse primers. The table shows the effectiveness of the new FAP primers in amplifying a broad spectrum of beta HPV types. Dots represent bases that are covered by the primers while the letters indicate the bases not covered by the primers, per each of the HPV types considered.

1.2 NGS data analysis: Characterization and taxonomic classification

The different PCR protocols described above were used to amplify randomly selected DNA extracted from skin swabs (n=119) and oral gargles (n=147) collected from healthy individuals (Table 8).

PCR pools	PCR protocols	Specimens	N	NGS pools
1	Beta-3-1	Skin swab	41	1
2	Beta-3-2		9	
3	FAP		52	2
4	FAPM1		54	3
5	CUT		57	4
6	FAPM2	Oral gargle	43	5
7	FAPM1		56	6
8	CUT		55	7
9	Beta-3-1		9	8
10	Beta-3-2		4	
11	FAP		11	
12	FAPM1		11	
13	FAPM2		12	

Table 8: Description of the PCR and NGS pools. Per each of the pools, the PCR protocol and the kind of specimen used are specified

After the amplification, the PCR products were mixed, generating eight pools (Table 8) and sequenced using Illumina MiSeq platform.

The NGS analysis produced a total of 50,017,076 paired-end raw reads that after quality trimming, de-replication, and chimeric PCR sequence removal, were reduced to 23,647,656 (47.3%). The PV related reads were identified by aligning them to the whole NCBI database (nr/nt, March 2017) using the MegaBlast algorithm. A total of 16,043,298 raw reads (roughly 67% of the filtered reads) were representative of PV sequences.

After, according to the RaxML-EPA classification, from the 119 skin samples, a total of 265 different PV types were identified (Figure 45 A; Table S1). The majority of the sequences were representative of the alpha (33.4%) and beta (29.5%) genera, describing the PV distribution in the skin.

Moreover, the 12.9% of the reads were assigned to taxonomically unclassified PV sequences (hereafter called “unclassified PVs”): bovine papillomavirus type 19 (BPV19), equine papillomavirus type 8 (EcPV8), Myotis ricketti papillomavirus 1 (MrPV1), Pudu puda papillomavirus type 1 (PpuPV1), and Sparus aurata papillomavirus type 1 (SaPV1). In addition, 9% of the reads were related to the genus gamma.

Using the FAPM1 protocol, a total of 107 PVs (8 alpha, 37 beta, 60 gamma, and 2 mu) were identified. CUT primers allowed the detection of 118 PVs (11 alpha, 36 beta, 68 gamma, 2 mu, and 1 nu) while FAP primers of 87 PVs (3 alpha, 34 beta, 49 gamma, and 1 mu).

The combined beta-3-1 and beta-3-2 protocols amplified mainly a non-human PV type: Colobus guereza monkey papillomavirus type 1 (CgPV1). Moreover, they allowed the detection of HPV16 (2 reads), five different beta HPV types (797,800 reads), of which three beta-3 types. The combination of the beta-3-1 and 2 protocols also allowed the detection of two non-referenced gamma HPV types (HPV-mDysk1 – KX781280 and HPV-mDysk6 – KX781285).

From the sequencing of 147 oral samples, a total of 161 different HPV types were identified. Here the most represented genus was the beta (29.5%), followed by gamma (19.6%) and alpha genera (7.8%) (Figure 45 B; Table S2).

Also, in the oral samples, a substantial fraction of reads were representative of taxonomically unclassified PV types: EcPV8, Miniopterus schreibersii papillomavirus type 1 (MscPV1), PpuPV1, and SaPV1 (Figure 45 B; Table S2).

In the oral samples, FAPM1 and FAPM2 protocols, allowed the detection of 55 different PV types (4 alpha, 30 beta, and 21 gamma) and 42 PVs (5 alpha, 21 beta, and 16 gamma), respectively.

The use of CUT protocol allowed the detection of 46 PV types (6 alpha, 17 beta, and 23 gamma) in the oral samples (Figure 45 B; Table S2).

A total of 745,860 reads in the skin and 163,448 reads in oral samples were assigned to taxonomically classified non-human PVs (i.e., PVs not belonging to the genera alpha, beta, gamma, mu, and nu) (Tables S1 and S2; Figure 45).

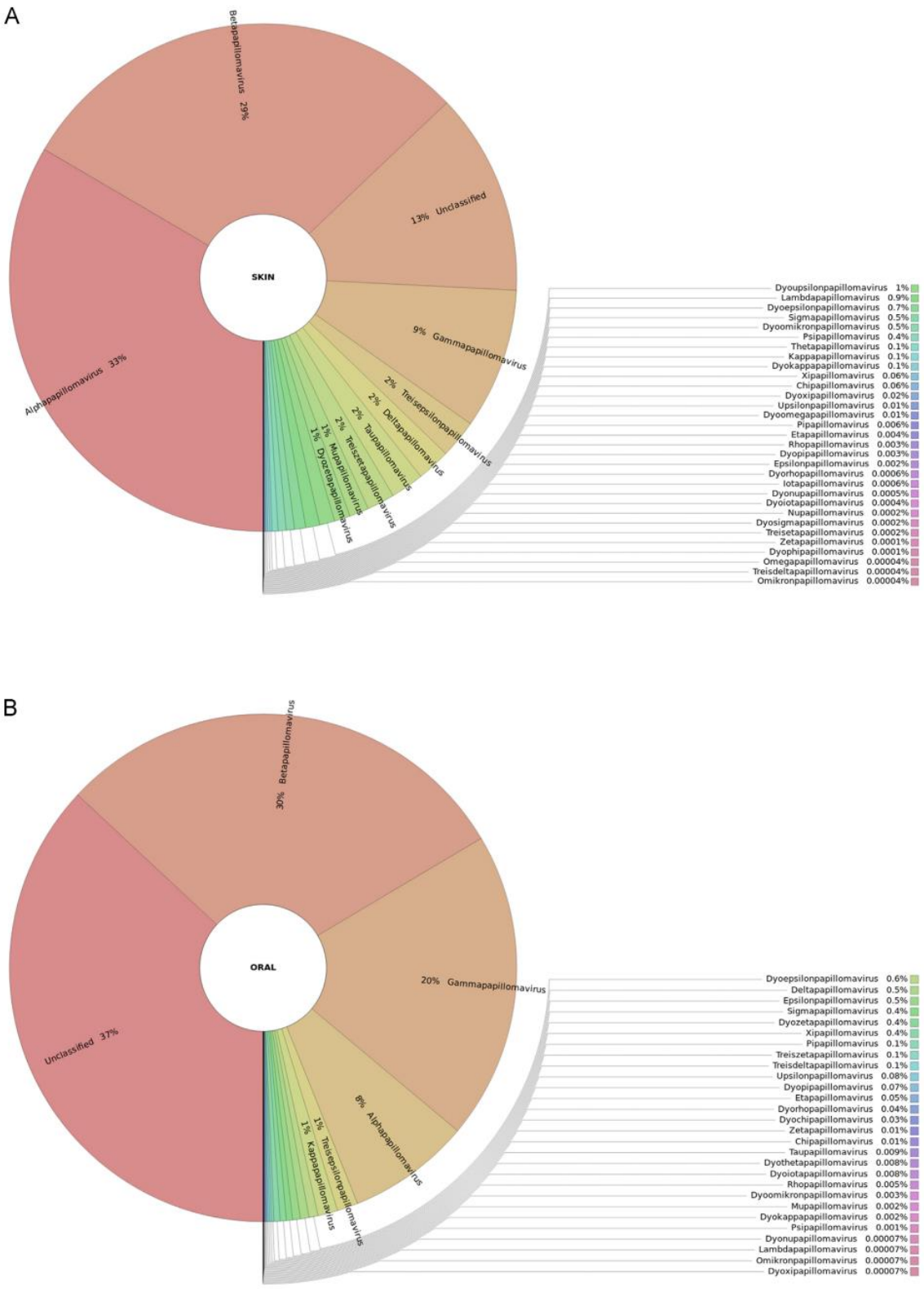


Figure 45: Absolute abundances (based on the number of reads) of different PV genera and species identified in this analysis: (A) skin samples, (B) oral samples

1.3 Subdivision of the NGS reads into known and putative novel PVs

Based on the initial blast analysis, performed using Megablast algorithm, the NGS sequences were divided into two groups: (i) known PV types (i.e., the sequences with ≥ 90% of similarity with the L1 genomic region of known PVs) and (ii) putative new PV types (i.e., the sequences with < 90% similarity with the L1 genomic region of any known PV type).

After this step of subgrouping, the sequences were subjected to the contigs generation step, using CAP3 de novo assembling tool, and then classified using RAxML-EPA tool.

A total of 8,002,617 reads were found to be representative of known PV types and the majority belonging to the genus beta (2,358,670 reads), followed by alpha (1,992,264 reads) and gamma (1,002,061 reads) (Figure 46 A).

A total of 1,678,061 reads was assigned to the “unclassified PVs” category, mainly represented by SaPV1 (KX643372.1).

A total of 2,588,649 reads (pool 1, Table 8), were produced using beta-3-1 and beta-3-2 primers in skin samples. Alpha was the most represented genus (56.6%) followed by genus beta (30.8%) (Figure 46 A).

Using the FAP primers (pool 2), a total of 985,675 reads were produced in skin samples, with the 40.8% representative of the genus beta, followed by genus gamma with the 14.5% of the reads (Table 8; Figure 46 A).

A total of 861,810 reads were generated in pool 3 using the FAPM1 protocol on skin samples.

The most represented genera detected using this protocol were alpha (23.3%), beta (13.6%), gamma (14.3%), and mu (7.6%), (Figure 46 A).

Using the same PCR protocol (FAPM1) in pool 6, a total of 244,587 reads were generated. Here, the beta genus was the most represented with 53.6% of the reads followed by genus gamma

(12.3%), while the alpha genus was poorly represented in this pool with the 0.1% of the reads (Table 7; Figure 46 A).

A total of 884,923 reads were generated using CUT primers on skin samples (pool 4). In pool 4, the most represented genus was gamma (23.3%) followed by beta (19.5%) and alpha (13%).

In oral samples (pool 7), CUT primers allowed the generation of 78.060 reads, and again the most represented genus was the gamma (17.2%) followed by beta (11%) and alpha (2.1%), (Table 7; Figure 46 A). In pool seven, a high proportion of reads (43.4%) were assigned to the category "unclassified PVs" with SaPV1 identified as the closest PV type to these sequences.

In pool 5, using FAPM2 protocol in oral samples, 466,004 reads were produced with a distribution of 9.3% alpha, 39.6% beta, and 32.5% gamma PV-related sequences (Table 7; Figure 46 A).

Additionally, in pool 8 were the products of five PCR protocols (Table 7) were pooled, a total of 1,892,909 reads were generated, of which 24.5% representative of beta, 8.8% alpha, and 17.6% gamma PVs. Also, in this pool, the highest proportion of reads (44.3%) was representative of "unclassified PVs" with SaPV1 identified as the closest PV type to these sequences. (Figure 46 and Table S2).

Considered altogether, the reads related to PV sequences identified in this analysis are representative of a total of 296 different known HPV types. These include: (i) 30 alpha types, of which 14 were found in skin samples, 8 in oral samples, and 8 in both tissues; (ii) 54 beta types of which 13 were from the skin, three from the oral cavity, and 38 from both tissues; (iii) 123 gamma types of which 70 were isolated from the skin, eight from the oral cavity, and 45 from both anatomical sites; (iv) 3 mu types of which one was found in the skin and 2 in both skin and oral samples; (v) 1 nu found in the skin.

Moreover, six unclassified PV types were detected, two in the skin, one in the oral samples and 3 in both sites (data not shown).

Regarding PV sequences not related to any of the main HPV genera (alpha, beta, gamma, mu, and nu), a total of 909,308 (11.3%) reads were identified, representative of 79 different PVs, of which 34 were identified in the skin, 11 in the oral cavity and 34 in both sites (data not shown).

A total of 19.032 reads identified, were representative of putative new HPV types (with < 90% similarity to known PVs).

The beta genus was the most represented in this group of sequences (35.6%) followed by gamma genus (23.2%), (Figure 46 B; Table S3).

By using beta-3-1 and beta-3-2 protocols in pool 1, 22 reads (26.8%) representative of 2 putative new beta-3-related sequences were identified (Fig. 3B; Table S3). Additional 54 reads (65.8%) were identified in the same pool, representing an unclassified PV. However, by considering a smaller contig from the same cluster, a correlation with Psipapillomavirus was found (Table S3). Six reads (7.3%) were related to Dyophipapillomavirus 1, but according to the PaVE classification, a correlation with HPV115 was found.

In pool two were the FAP protocol was used on skin samples, 40 reads (1.2%) were representative of new beta HPV types, and 2228 reads (69.2%) of new gamma types.

A total of 116 reads were representative of unclassified PV types according to the RAxML-EPA classification. Among them, two sequences were related to MTS2[623] (gamma-7) according to the PaVE classification.

Additional 833 reads were assigned to Taupapillomavirus 3, 4 reads to Deltapapillomavirus 5, and 3 reads to Dyorhopapillomavirus 1 (Table S3).

In pool 3, where FAPM1 was used on skin samples, 294 reads (70.2%) were representative of new beta, 48 reads (11.5%) of new gamma, 52 reads of new delta-2 (12.4%), and 7 reads of new lambda-3 (1.7%) PV types.

In pool 6, where the same protocol was used with oral samples, 21 reads were assigned of putative new PVs, of which 9.5% related to beta-1 HPV types, 23.8% to Sigmapapillomavirus 1, and 66.7% to Dyoiotapapillomavirus 2, considering the RAxML-EPA (Figure 46 B; Table S3).

In pool 6, no putative new gamma types were identified.

In pool 4, the use of CUT protocol on skin samples allowed the identification of 2126 reads (72.7%) representative of putative new gamma HPV types and a smaller fraction (4.2%) of beta types. In the same pool, 12 reads were related to nu, 12 to alpha types (species alpha-2 and alpha-3), and two reads to Lambdapapillomavirus 3 (alternatively identified as canine papillomavirus 6 according to the PaVe classification). An additional eight non-human PVs were identified (Table S3) in the same pool.

In pool 5, when FAPM2 was used on oral samples, 6295 reads (99.9%) were identified as putative new beta, and the 0.1% of the reads were representative of putative new gamma types (Figure 46 B; Table S3).

In pool 7, when the CUT protocol was used with oral samples, a putative new non-human PV, related to Chipapapillomavirus 2, and an “unclassified type” were identified according to the RAxML-EPA classification.

Both these two sequences were assigned to species beta-1 when the PaVe classification was considered (Table S3).

For pool 8, the majority of reads were assigned to unclassified PVs (3308 reads), Treisdeltapapillomavirus 1 (2713 reads), and other non-human PV genera (Treisepsilonpapillomavirus, Treisdeltapapillomavirus, and Treiszetapapilloamvirus; 16 reads). In this pool, a minority of reads were assigned to beta (0.07%) and mu (0.3%) HPVs.

In total, 105 putative novel HPV types were identified, including 29 beta HPVs, of which 21 were found in skin and 8 in oral samples.

Regarding gamma genus, 30 putative novel HPV types were identified in skin and 2 in oral samples, while for the alpha genus, only two putative new viruses were identified in the skin.

One putative new mu HPV type was identified in the skin, and 24 PVs, not belonging to any known HPV genera, were identified, of which 17 were found in the skin and 7 in oral samples.

Finally, 15 unclassified PVs were identified in the skin and 2 in the oral samples. Nine of these reads were found to be related to beta and eight to gamma HPV types, using the PaVe classification.

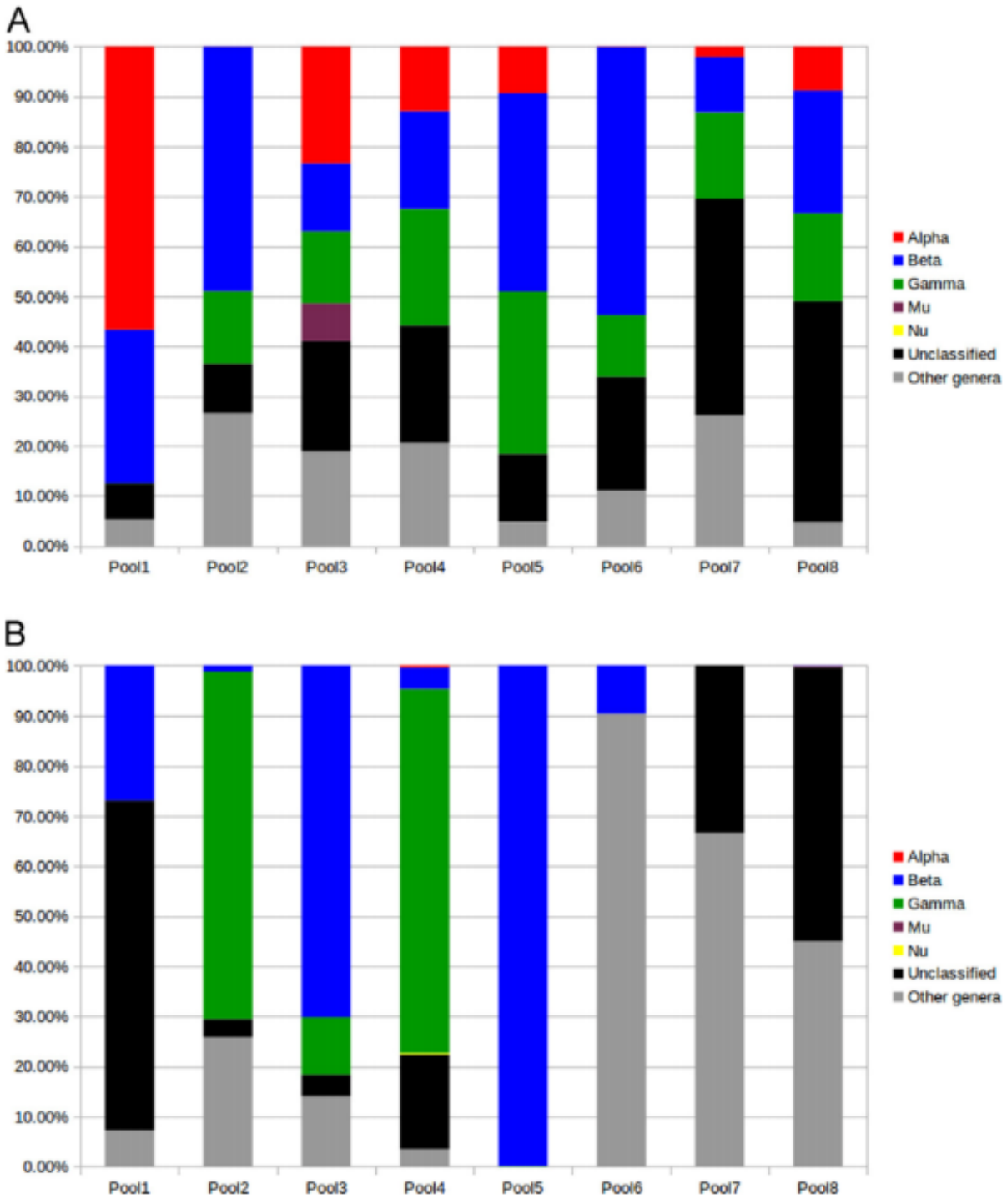


Figure 46: (A) Percentage of reads, representative of the different known PVs detected in each NGS pool; (B) Percentage of reads, representative of the different putative new PVs detected in each NGS pool (RAxML-EPA)

2. Pipeline improvements and development of PVAmpliconFinder

In the second step of our work, the strategy used to analyze the NGS sequencing data for the detection of putative novel HPV types was further implemented and automatized. PVAmpliconFinder, a data analysis workflow designed to rapidly identify and classify known and potentially new Papillomaviridae sequences from NGS amplicon sequencing with degenerate PV primers, was developed.

Published: Alexis Robitaille, **Rosario N Brancaccio**, Sankhadeep Dutta, Dana E. Rollison, Marcis Leja, Nicole Fischer, Adam Grundhoff, Tarik Gheit, Massimo Tommasino, Magali Olivier. PVAmpliconFinder: a workflow for the identification of human papillomaviruses from high throughput amplicon sequencing. NAR Genomics and Bioinformatics. 2020

2.1 NGS and data processing

The PVAmpliconFinder workflow was applied to the data obtained from a preliminary study where targeted sequencing NGS was used to identify new PVs in human skin swab specimens and oral rinses from healthy individuals.

The samples were amplified using different primers, all targeting a portion of the HPV L1 ORF.

After the amplification, the samples were grouped based on the PCR protocol used, and eight pools were generated, and paired-end sequenced using the Illumina MiSeq system. A total of 2.65 million raw reads were generated (331,359 raw reads on average per sample pool) (Table 9 A). After the quality trimming, around 2% of reads were removed (Table 9 A).

The merging of paired reads reduced by at least two-fold the total number of sequences and extended the average size of the sequences. Around the 90% of reads were merged for all the samples, except for DNA sample pool 6, where only 60% of the reads were merged (Table 9 A). FASTQC report enabled the identification of primer contamination in about 10% of the reads

explaining this low level of merged reads for pool 6. After the de-replication step, the number of sequences was dramatically reduced, showing that the different amplicons were highly represented (Table 9 B). Less than 1% of the sequences were identified as potentially chimeric (Table 9 B). After, the clustering step further reduced the number of unique sequences retained, from about 8% to 1% of the total sequences considered in the upstream step (Table 9 B). In total, about 28.5% (756,506/2,650,877) of the raw reads were retained for the MegaBlast step (Table 9 B). More than 90% of the centroid-clustered unique sequences of the five pools from skin swab specimens (S1-S5) had their best match in a Papillomaviridae family sequence (Table 9 B). Regarding the three pools from oral rinses (S6-S8), about 86.5% of the centroid-clustered unique sequences were representative of Papillomaviridae family sequences (Table 9 B).

A

Step	Total sequencing raw reads		TrimGalore		Merging		Dereplication		Chimeric identification		Clustering	
Samples	N paired-end reads		N paired-end reads	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%
S1	564435		564064	99.93	551266	97.73	22551	4.09	22498	99.76	79	0.35
S2	62148		61708	99.29	58031	94.04	3281	5.65	3268	99.60	162	4.96
S3	316297		315999	99.91	307400	97.28	15562	5.06	15562	100.00	51	0.33
S4	109441		109326	99.89	106406	97.33	4842	4.55	4822	99.59	62	1.29
S5	309779		309390	99.87	294563	95.21	14101	4.79	14091	99.93	140	0.99
S6	554415		551742	99.52	331648	60.11	13820	4.17	13738	99.41	1162	8.46
S7	470655		467944	99.42	421764	90.13	28729	6.81	28659	99.76	609	2.12
S8	263707		263270	99.83	244177	92.75	13293	5.44	13283	99.92	194	1.46
Total num	2650877		2643443	99.72	2315255	87.58	116179	5.02	115921	99.78	2459	2.12

B

Step	Dereplication		Chimeric identification		Clustering		Papillomaviridae best hit (eval<=1e-5)		Defined group (same best hit)					
									Putative new (>10% dissimilarity)		Putative known (<10% dissimilarity)			
Samples	N sequences	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%	N sequences	%		
S1	22551	4.09	22498	99.76	79	0.35	61	77.22	0	0	5	8.20		
S2	3281	5.65	3268	99.60	162	4.96	138	85.19	0	0	6	4.35		
S3	15562	5.06	15562	100.00	51	0.33	49	96.08	1	2.04	18	36.73		
S4	4842	4.55	4822	99.59	62	1.29	62	100.00	0	0	28	45.16		
S5	14101	4.79	14091	99.93	140	0.99	129	92.14	2	1.55	39	30.23		
S6	13820	4.17	13738	99.41	1162	8.46	910	78.31	0	0	16	1.76		
S7	28729	6.81	28659	99.76	609	2.12	513	84.24	0	0	16	3.12		
S8	13293	5.44	13283	99.92	194	1.46	188	96.91	0	0	8	4.26		
Total num	116179	5.02	115921	99.78	2459	2.12	2050	83.37	3	0.45	136	16.73		

Table 9: Number of sequences at each step of the workflow. A) From the raw data to the clustering step. B) From the dereplication step to the end of the analysis

2.2 Taxonomic classification of the PV-related sequences

In DNA sample pool S5, two putative new PV sequences represented by five reads and 39 putative known PV sequences represented by 60,892 reads were detected. One of the putative new PV sequences in this pool was represented by three reads (PV_2). The MegaBlast algorithm (using the full "nt" database) aligned it against "Gammapapillomavirus 13 isolate Gamma13_HIVGc158, complete genome" (MF588722.1) with 81.25% of identity. The Gamma13_HIVGc158 is a complete genome not reported in the Papillomavirus Episteme (PaVE) database. The BlastN algorithm (using the PaVE database) aligned this sequence against HPV-mEV03c45 (MF588721), an unreferenced Gamma PV genome, with 78.69% of identity. RaxML-EPA placed this sequence in the reference tree close to HPV213 (MF509818), also a potential Gamma PV. The MegaBlast algorithm identified the best hit for this sequence in a partial cds (342 bp) of a major capsid protein L1 gene (isolate GC12_1; FJ969907.1) with nearly 99% of identity. The BlastN alignment against the PaVE database found the best hit in a Gamma-10 referenced PV genome (HPV130; GU117630), with a percentage of identity of the 86.12%. Finally, RaxML-EPA classification found EdPV2 (MH376689) as the closer PV sequence. PV_2 may represent a novel PV type, but a full characterization is required for the correct classification.

2.3 The relative unnormalized abundance of PV sequences

The relative unnormalized abundance of PV-related sequences identified by MegaBlast, BlastN, and RaxML-EPA is represented in Figures 47 A, B, and C, respectively. Detailed taxonomic assignations were reported in supplementary tables based on MegaBlast, BlastN, and RaxML-EPA, respectively (Data not reported in this thesis). Beta-3 species constitutes the most represented species, with 42% of beta-3 related sequences identified by MegaBlast, and 62%

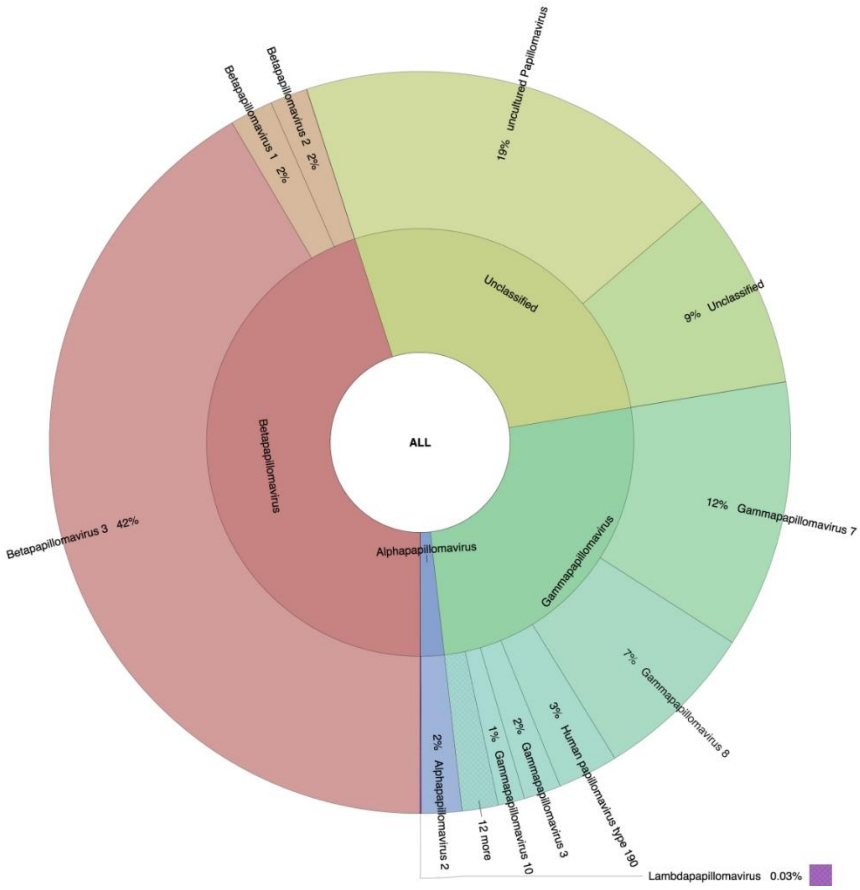
identified by both BlastN and RaxML-EPA (Data not showed). The second most represented group was constituted by “unclassified” sequences identified by MegaBlast (28% of the sequences), using the NCBI nt database.

Based on MegaBlast identification, the third most represented genus was the gamma genus (i.e., 24% of gamma-related sequences), followed by the alpha genus (2%) and a small proportion of Lambdapapillomavirus (0.03%) due to the identification of a feline PV partial cds sequence (EF535004.1) in sample pools 1 and 2 (Data not shown).

Based on BlastN and RaxML-EPA identification, the second most represented group was constituted by unreferenced PVs, with a major subset putatively classified as unreferenced Gammapapillomavirus sequences (about 17%) and a small subset as unreferenced Betapapillomavirus sequences (about 1%).

According to both BlastN and RaxML-EPA identification, the third and fourth most represented genera were the referenced gamma and alpha PVs (Figure 47 B and C). BlastN could not classify 0.008% of the sequences due to the best subject sequence associated with an e-value under the threshold defined as 1e-1 (Data not showed). RaxML-EPA classified the 0.8% of the sequences as “unclassified”. Those sequences presented homology to a newly described Erethizon dorsatum PV (EdPV2; MH376689). Forty-six reads that were unclassified by BlastN (due to the e-value threshold) were classified as Taupapillomavirus by RaxML-EPA, with homology to Felis catus PV type 4 and 5 (Data not shown).

A


ALL
Betapapillomavirus
Unclassified
Gammapapillomavirus
Alphapapillomavirus
Betapapillomavirus 3 42%
Betapapillomavirus 1 2%
Betapapillomavirus 2 2%
19% uncultured Papillomavirus
9% Unclassified
12% Gammapapillomavirus 7
7% Gammapapillomavirus 8
3% Human papillomavirus type 190
2% Gammapapillomavirus 3
1% Gammapapillomavirus 10
12 more
2% Alphapapillomavirus 2
Lambdapapillomavirus 0.03%

B


ALL
Beta
Alpha
Gamma
Unreferenced
Betapapillomavirus 3 62%
4% Betapapillomavirus 2
2% Betapapillomavirus 1
17% Gammapapillomavirus
1% Betapapillomavirus
7% Gammapapillomavirus 8
2% Gammapapillomavirus 10
2% Gammapapillomavirus 3
0.7% Gammapapillomavirus 7
9 more
2% Alphapapillomavirus 2
Unclassified 0.008%

C

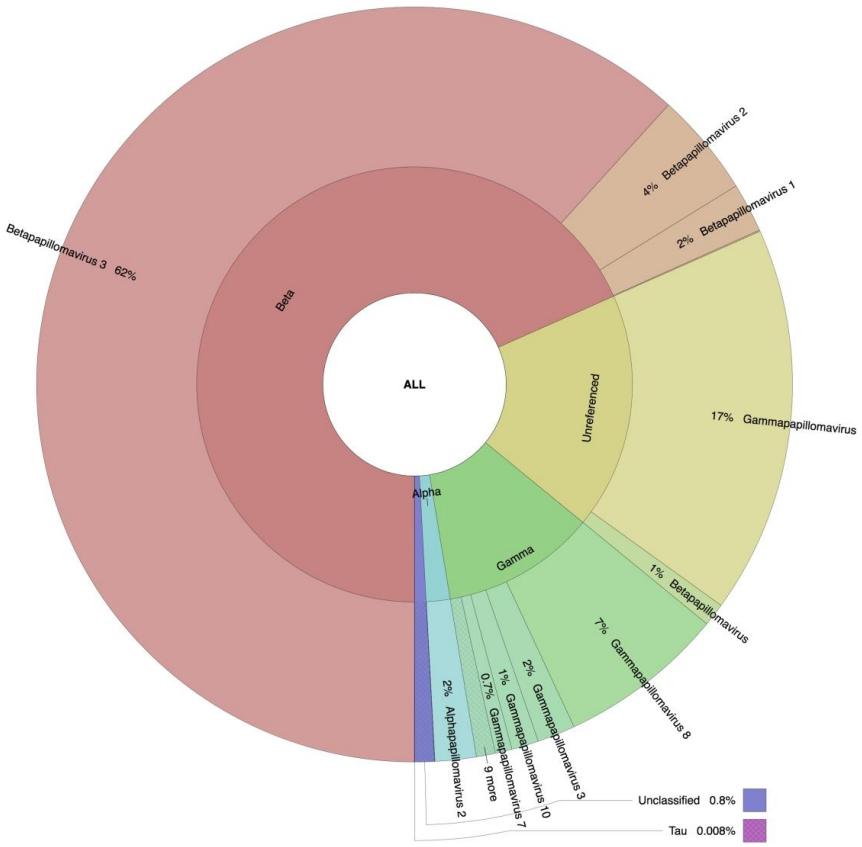


Figure 47: Graphical representation of the unnormalized abundance of PV genera and species in terms of number of reads based on (A) Megablast, (B) BlastN and (C) RaxML-EPA

2.4 Discovery and characterization of putative new PV-related sequences

A total of three putative novel papillomaviruses were identified in this study, named PV_1, PV_2 and PV_3 (Table 10). According to MegaBlast, the best blast hit for PV_1 is an unreferenced Gamma-12 complete genome. Based on the BlastN and RaxMLEPA, PV_1 has its best hit with HPVmSK197 (MH777339), from the PaVe database. MegaBlast find the best hit for PV_2 in an unreferenced Gamma-13 complete genome (MF588722). Using the BlastN classification, PV_2 found its best hit in HPV-mEV03c45 (MF588721), an unreferenced Gammapapillomavirus genome, and RaxML-EPA to HPV213 (MF509818), a referenced, but unofficially classified Gammapapillomavirus genome.

For PV_3, MegaBlast found the best hit in an unclassified partial cds of the isolate GC04 (FJ969896), while it was related to the unreferenced HPV-mSK014 (MH777162) by both BlastN and RaxML-EPA identifications. The sequences representative of these putative new PVs ranged from 160 to 372 nucleotides in size, and all sequences presented more than 15% of dissimilarity with non-referenced PV sequences based on MegaBlast. All were amplified from skin DNA samples, using FAP[592] and CUT[383] primers.

A previous analysis of the same data had led to the characterization of the full genome sequence of a novel Gamma-8 PV (Table 10, "37VIRUSput")[624]. The official name HPV224 was assigned to this new virus. PVAmpliconFinder performed a correct classification and identification of this new virus.

VIRUSname	PV_1	PV_2	PV_3	37VIRUSput
%dissimilarity	16.67	18.75	20.49	0.85
Abundance	0.0025	0.0049	0.0033	3.2918
N°reads	2	3	2	698
GInum	gi|1273499301|gb|MF588716.1|	gi|1273499348|gb|MF588722.1|	gi|270048224|gb|FJ969896.1|	gi|1214938671|gb|MF356498.1|
AlignmentPosition_MegaBlast	3-78	1-352	207-327	1-236
VIRUS_closest_MegaBlast	Gamma12_EV07c367 isolate,c.g.	Gamma13_HIVGc158 isolate, c.g.	HPV isolate GC04 L1 gene,	HPV isolate ICB1, c.g.
Pool	pool3-skin-pathogen_S3_L001	pool5-skin-pathogen_S5_L001	pool5-skin-pathogen_S5_L001 partial cds	pool4-skin-pathogen_S4_L001
Tissu	skin	skin	skin	skin
Primer	FAP	CUT	CUT	FAPM1
Length	160	353	372	262
Align.Pos-BlastN_start:stop(length)	1-160(160)	1-352(352)	3-370(368)	1-255(255)
VIRUS_closest_Blast	HPV-mSK197(92.5%)	HPV-mEV03c45(78.69%)	HPV-mSK014(94.57%)	HPV224(97.25%)
BlastN_Classification	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus
RaxML_closest_PV	HPV-mSK197	HPV213	HPV-mSK014	HPV224
RaxML_Classification	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus	Gammapapillomavirus

Table 10: Putative new PV-related sequences identified using PVAmpliconFinder

Specific aim #2

1. Isolation of a novel beta-2 human papillomavirus from skin

Starting from the NGS data, described in the paragraph “Isolation of novel HPV types using broad spectrum primers and NGS” of “specific aim #1” of the present work, a novel beta-HPV type was identified. After, using Sanger sequencing and a primer walking strategy, its genome was fully characterised.

Published: **Brancaccio RN**, Robitaille A, Dutta S, Rollison DE, Tommasino M, Gheit T. Isolation of a Novel Beta-2 Human Papillomavirus from Skin. Microbiol Resour Announc. 2019 Feb 28; 8(9):e01628-18g

1.1 Full characterization of the novel HPV genome

Starting from the NGS data described in the aim #1 of the present study, a partial L1 region sequence (99 bp) representing a putative new beta HPV type was isolated from the skin samples.

Specific primers were designed based on this partial L1 sequence and used to screen the original skin samples to identify the one positive for the specific putative new virus.

After the identification of the positive sample, rolling circle amplification (RCA) was performed to enrich the sample of viral copies, according to the manufacturer’s instructions (illustra TempliPhi 100 amplification kit; GE Healthcare, USA).

To obtain the whole HPV genome, long-range PCR was performed on the RCA product using PrimeSTAR GXL DNA polymerase (TaKaRa Bio), outward-directed primers specific for HPV ICB2 (forward primer, 5′-CAGACAGAACACATCTTTTGATCC-3′; and reverse primer, 5′-TCGTCCCGTGACCCACCCTGA-3′).

An amplicon of roughly 8 kb was generated and cloned in pCR-XL-2 TOPO vector using the TOPO XL-2 complete PCR cloning kit, following the manufacturer's instructions (Invitrogen, Carlsbad, CA).

Once the complete genome of the virus was cloned into a vector, a primer-walking strategy was used to obtain the whole sequence of the virus, by performing consecutive Sanger-sequencing runs (GATC Biotech, Germany).

The complete genome was sequenced at least twice to correct potential sequencing errors.

A total of 31 sequences were produced in the analysis and aligned to reconstruct the whole genome. CAP3[625] sequence assembly program was used, with default parameters, to assemble these sequences, thus obtaining the entire genome of the virus.

The clone of the virus was submitted to the International Human Papillomavirus Reference Center in Stockholm (www.hpvcenter.se) for assignment of HPV type number, and the official name HPV227 was assigned to this new virus.

The analysis of the L1 ORF nucleotide sequence revealed an identity of the 87.9% with HPV37, identified as the closest HPV type. Being this percentage of identity less than 90%, defines HPV227 as a novel beta-2 HPV type[170].

The G+C content of ICB2 is 40.7%. The virus has the typical genome organization of other cutaneotrophic HPVs, with five early (E1, E2, E4, E6, and E7) and two late (L1 and L2) ORFs, and no E5.

The long control region (LCR) is 382 bp and contains two polyadenylation sites (AATAAA) for L1 and L2 transcripts and four consensus palindromic E2-binding sites, as follows: ACCG-N_4-CGGT (*n = 2*), ACC-N_5-GGT (*n = 1*), and ACC-N_1-GGT (*n = 1*). A putative TATA box domain (TATAAGA) for the downstream early promoter is also present.

The two conserved zinc-binding domains of the viral E6 protein [CxxC$(x)_{29}$CxxC and CxxC$(x)_{30}$CxxC] are present and are separated by 36 amino acids[183].

In the E7 protein are present one zinc-binding domain [CxxC$(x)_{29}$CxxC] and one LxCxE motif[183]. In the carboxy terminus of the E1 protein was identified one ATP-binding site (GPPDTGKS) for ATP-dependent helicase activity.

The complete genome sequence of HPV ICB2 is available in GenBank under accession number MK080568.

Specific aim #3

1. MinION sequencing for the reconstruction of the whole genome of HPV ICB2

The genome of HPV ICB2 was characterized in specific aim #2 of the present work, using a primer walking strategy. Here, the MinION sequencing technology (ONT) was tested to evaluate his capability in sequencing the whole genome of HPV ICB2. A new bioinformatics tool and a pipeline were developed for the analysis of the MinION sequencing data.

Under review: **Rosario N. Brancaccio**, Alexis Robitaille, Sankhadeep Dutta, Dana E. Rollison, Massimo Tommasino, Tarik Gheit. MinION nanopore sequencing and assembly of a complete human papillomavirus genome. Virology Journal

1.1 MinION sequencing and assembly using three independent runs (Protocol A)

The 1D PCR barcoding amplicons (SQK-LSK109) protocol was used to generate three different DNA libraries, after sequenced using a single MinION flow cell (R9.4.1), performing three consecutive 12-hours sequencing runs. The protocol “check your flow cell” in the MinKNOW software, was used to measure the number of available nanopores before each sequencing run.

The number of active nanopores decreased from run 1 to 3 (1577, 1124, and 857 respectively), maintaining the guaranteed level for an optimal sequencing run (800 pores). In total the runs generated 9,354,933 raw reads (run 1: 3,186,245 reads; run 2: 2,464,705 reads; and run 3: 3,703,983 reads). Roughly 92% of the reads generated by runs 1 and 2, and 82.8% of the reads from run 3 passed the quality control (QC) filtering (Table 11).

Protocol	Run	Active pores (N)	Running time (hrs)	Barcodes used	Reads with barcode (%)	Raw reads (N)	Giga-bases called (Gb)	QC filtered reads (%)	Mean read length (nt)	N50 (nt)	Mean ICB2 coverage (fold)	Mean similarity to HPV-ICB2 (%)
A	1	1577	12	BC01	85.9	3,186,245	6.44	92.7	1968	7329	545289	95.2
	2	1124	12	BC02	78.0	2,464,705	5.61	92.0	2288	7391	443143	95.7
	3	857	12	BC03	75.3	3,703,983	4.05	82.8	1032	7404	227763	95.4
B	4	560	48	BC01/02/03	66.2	3,255,879	4.05	70.1	1507	7362	200454	94.9

Table 11: Sequencing output metrics for the four MinION sequencing runs

All the first three runs have an N50 greater than 7300 nt (Table 11). The cumulative number of reads sequenced throughout the sequencing time are shown in Figure 48.

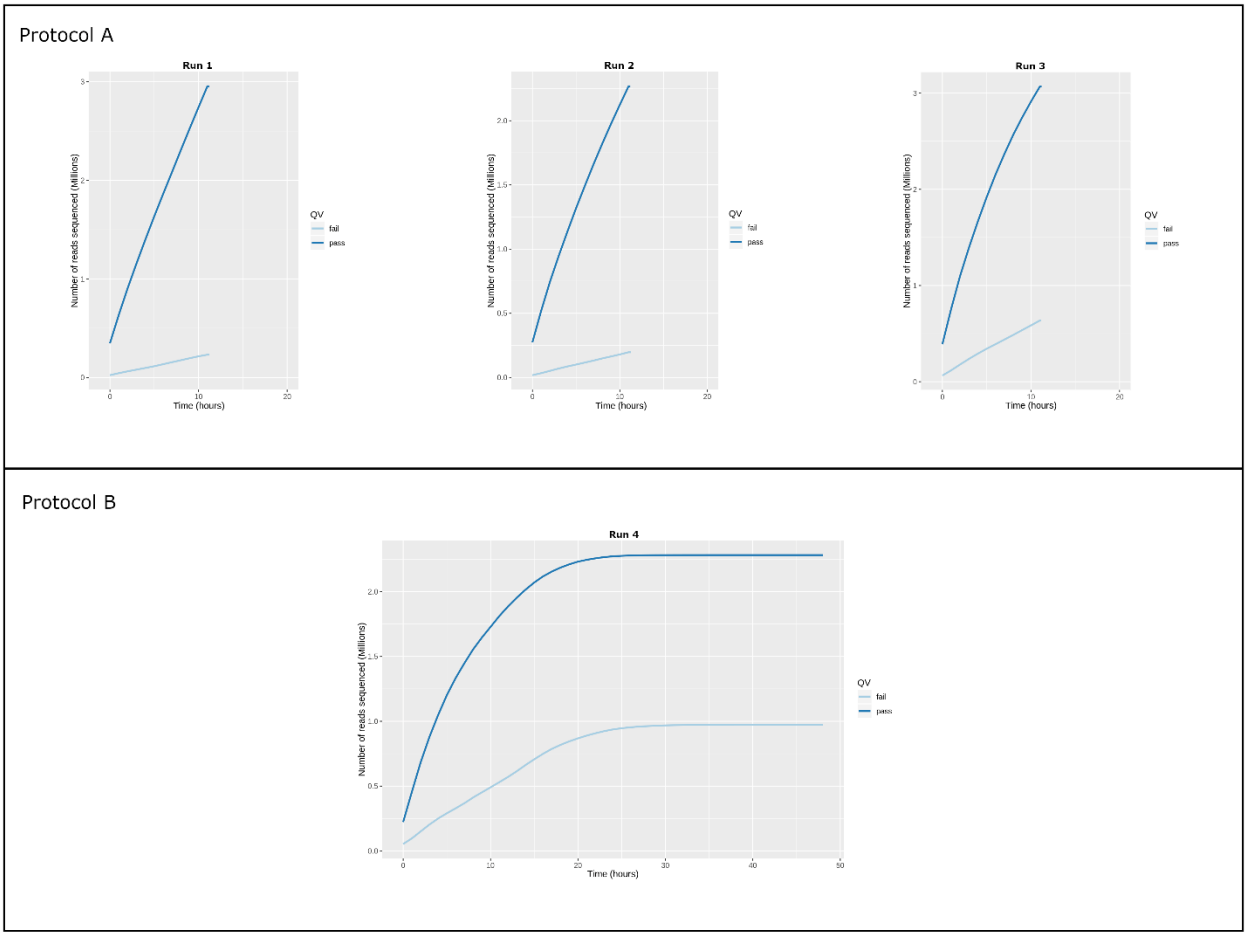


Figure 48: Number of reads generated. Three 12-hour runs (run 1-3; Protocol A), and a 48-hour single run (run 4) with pooled libraries (Protocol B) were performed using a FLO-MIN106D flow cell.

The blue line represents the number of reads passing the QC filtering, while the light blue line represents the number of reads that did not pass the filtering step.

From run 1 to 3, there was an increase in the proportion of reads passing the QC filtering from 7.3% to 17.2%. The coverage of the HPV-ICB2 genome was higher in the first run (mean coverage: 545,289-fold) compared to that observed in runs 2 and 3, which had coverages of 443,143-fold and 227,763-fold, respectively (Table 11 and Figure 49 A).

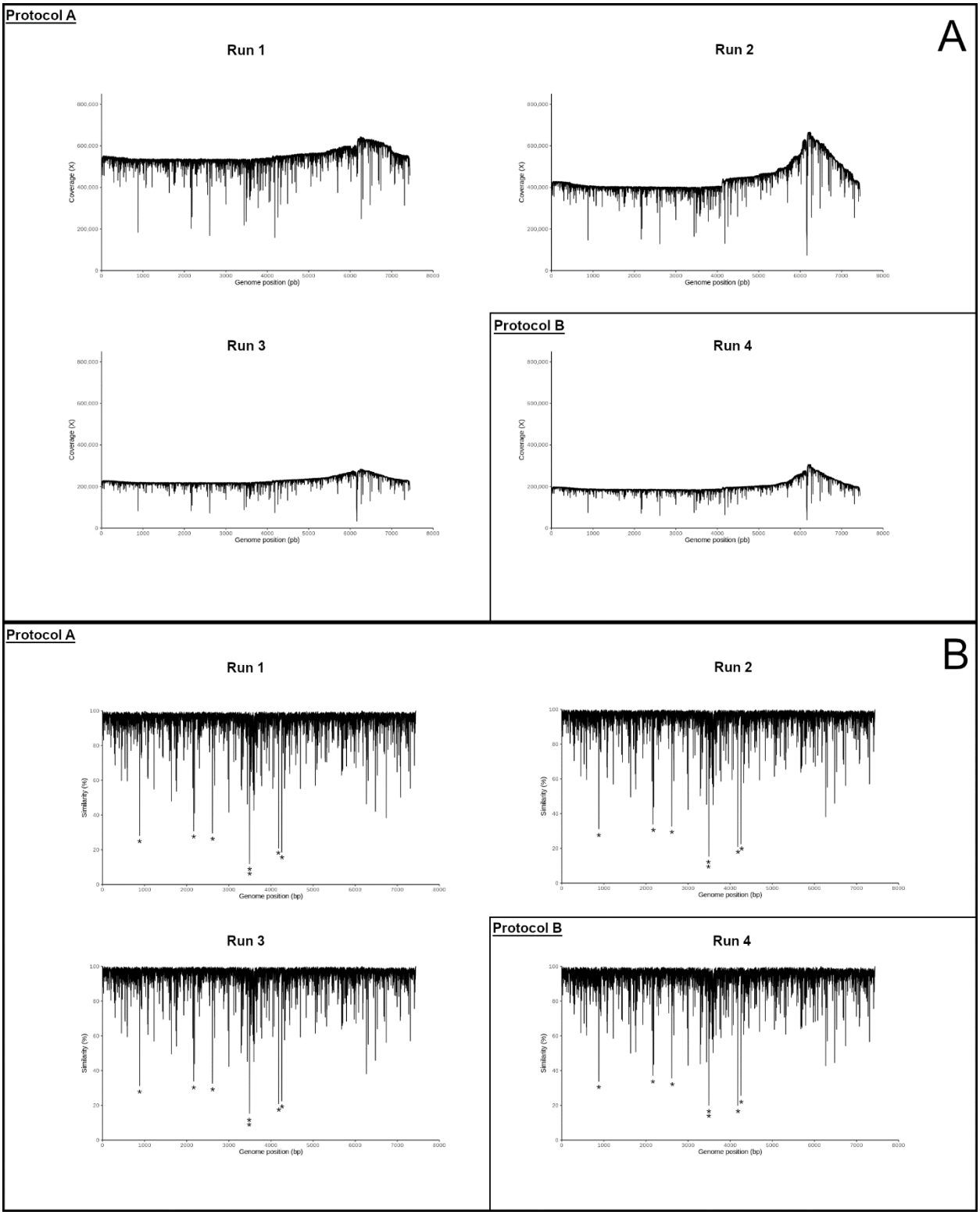


Figure 49: (A) MinION reads coverage calculated using HPV-ICB2 genome as reference. (B) Percentage of similarity among each base when aligned to the HPV-ICB2 reference genome. An asterisk indicates the eight nucleotide positions with the lowest similarity percentage to HPV-ICB2 among the four runs.

Analysis of the similarity between the sequencing reads and the reference genome of HPV-ICB2 (Figure 49 B) showed the presence of seven nucleotide positions (positions 882, 2165, 2612, 3491, 3492, 4182, and 4256) with the highest error rate in all three runs.

The average error rates were 4.8, 4.3, and 4.6% for runs 1, 2, and 3, respectively, considering the raw sequences (Table 11).

BLAST tool was used with the Megablast algorithm and the whole nucleotide collection (nr/nt, February 2019) database, to identify the MinION sequences. The majority of the reads found their best match in HPV-ICB2 (78.7, 90.8, and 84.0% for runs 1, 2, and 3, respectively). The remaining reads were representative of bacteria, archaea, eukaryotes, cloning vector sequences, and other viral sequences (Table 11).

After read filtering and SPAdes assembly, a second BLAST alignment was performed using the BLASTn algorithm (Figure 43), and thus 99.9% of the contigs where identified as HPV-ICB2 in all three runs (data not shown).

After, the contigs generated by CAP3 were used to reconstruct the whole genome of HPV-ICB2. For runs 1, 2, and 3, the percentage of identity to HPV-ICB2 was 97.56, 97.66, and 99.02%, respectively. Following, two polishing steps, using the DraftPolisher and Nanopolish tools, were performed (Figure 43), improving the percentage of identity with HPV-ICB2 to 99.95, 99.93, and 99.93% for runs 1, 2 and 3, respectively.

Lastly, after the alignment of the three genomes using MUSCLE, a consensus sequence was generated. The final percentage of identity to HPV-ICB2 was 99.93%, including five nucleotide gaps. The single nucleotide gaps were placed within homopolymeric A (*N*=1), T (*N*=2), C (*N*=1), and G (*N*=1) stretches at nucleotide positions 882, 2165, 2612, 4182, and 6268. Sanger sequencing was performed, verifying that those nucleotide gaps were not present in the DNA that was sequenced using the MinION.

As already reported [626,627], we observed the presence of 193,850 BC1 reads in run 2 (9.2% of the barcoded reads), and 27,343 BC1 and 116,625 BC2 reads in run 3 (4.9% of the barcoded reads) due to carry-over DNA from previous runs, despite intermediate washing steps.

1.2 MinION sequencing and assembly using a single run (Protocol B)

A final 48-hour run (run 4) was performed with BC01, BC02, and BC03 barcoded DNA libraries mixed in a single pool, using the remaining 560 active nanopores of the flow cell. The run was performed until the complete exhaustion of the nanopores. A total of 3,255,879 reads were generated in this run, among which 70.1% passed the QC filtering, and with an N50 value of 7362bp. Using BLAST, more than 90% of the reads generated by run 4 found their best match in HPV-ICB2, with a mean coverage of 200,454-fold (Table 11 and Figure 49 A). The similarity analysis of the MinION reads to the reference HPV-ICB2 genome showed that the seven nucleotides with the highest error rate were placed at the same nucleotide positions observed in runs 1-3 (Figure 49 B). More than 99% of the contigs were identified as belonging to the HPV-ICB2 genome after reads filtering and SPAdes assembly (data not showed).

The assembly of the contigs generated by CAP3 from each barcoded library allowed the reconstruction of the entire genome of HPV-ICB2, with percentages of identity to the reference genome of HPV-ICB2 of the 99.34, 98.69, and 96.86% for BC1, BC2 and BC3, which increased to 99.89, 99.87, and 99.92% after two polishing steps, respectively.

Finally, after the alignment of the three genomes performed with MUSCLE, a consensus sequence was obtained. The final percentage of identity to the reference genome of HPV-ICB2 was 99.89% (8 nucleotide gaps).

All the single nucleotide gaps identified, were located within homopolymeric A (*N*=1), G (*N*=4), and C (*N*=3) stretches at nucleotide positions 2197, 3350, 3491, 3513, 3574, 3583, 3586, and 6025. Sanger sequencing was performed on the DNA libraries loaded into the flow cell to confirm the absence of those gaps before the sequencing run.

1.3 Effect of run time on final assembly quality

MinION sequencing data were analyzed based on cumulative run-time to simulate the effect of run-length on final assembly quality. The analysis was performed on reads generated in the first 3, 6, 9, and 12 hours. Three hours of run time were sufficient to assemble the entire viral genome. After 3 hours, the coverages ranged from 15,973-fold to 138,219-fold, and the mean similarities ranged from 94.9 % to 95.9%, allowing the reconstruction of the whole HPV-ICB2 genome with percentages of identity, against the reference genome of the virus, exceeding 99.9% for both protocols (Table 12, and data not showed).

Protocol	Run	Library ID	Running time (hrs)	Mean HPV-ICB2 coverage (fold)	Mean Similarity to HPV-ICB2 (%)	Barcoded Reads (N)	N50 (nt)		Mean read length (nt)	Identity to HPV-ICB2 (%)
A	1	BC01	3h	138219	95.2	842252	7302		1634	99.97
			6h	283704	95.2	1514178	7315		1836	-
			9h	418898	95.2	2137181	7323		1924	-
			12h (end of the run)	545289	95.2	2737785	7329		1968	99.95
	2	BC02	3h	116642	95.9	608078	7347		1874	99.96
			6h	237855	95.8	1101879	7369		2099	-
			9h	348430	95.8	1536247	7382		2219	-
			12h (end of the run)	443143	95.7	1922810	7391		2288	99.93
	3	BC03	3h	68785	95.7	990975	7350		828	99.91
			6h	136892	95.6	1742726	7381		943	-
			9h	190719	95.6	2334682	7396		1001	-
			12h (end of the run)	227763	95.4	2789845	7404		1032	99.93
B	4	BC01	3h	15973	95.5	136650	7358		1344	99.95
			6h	32392	95.5	253906	7385		1501	-
			9h	44026	95.4	336064	7401		1596	-
			12h	51420	95.4	396188	7406		1640	99.91
			48h (end of the run)	56462	95.0	498718	7393		1533	99.89
		BC02	3h	23237	95.4	142496	7349		1867	99.95
			6h	46827	95.4	266743	7381		2061	-
			9h	63568	95.3	356610	7397		2173	-
			12h	74176	95.3	423495	7402		2220	99.95
			48h (end of the run)	81582	94.9	538160	7389		2067	99.87
		BC03	3h	23812	94.9	390006	7261		879	99.93
			6h	40835	94.9	652878	7306		919	-
			9h	51855	94.8	818088	7331		957	-
			12h	58344	94.8	930999	7335		974	99.92
			48h (end of the run)	62410	94.7	1117242	7303		923	99.92

Table 12: Effect of run time on the sequencing output metrics and sequencing quality data

Discussion

Specific aim #1

1. Identification of novel HPV types using a targeted NGS approach

The first HPV type was discovered in 1978 by Orth and colleagues[628], and to date, according to the Papillomavirus Episteme database (https://pave.niaid.nih.gov/)[172], 66 Alphapapillomaviruses, 67 Betapapillomaviruses, 301 Gammapapillomaviruses, 5 Mupapillomaviruses, and 1 Nupapillomavirus, have been discovered (update May 2020).

In the past, broad-spectrum primers, such as FAP and CUT primers, have been used successfully to identify new HPV types[382,383].

Though, this strategy is quite laborious and time-consuming, enabling mainly the identification of the most represented HPVs and resulting ineffective in the context of multiple infections.

With the advent of NGS and other molecular biology tools, in the last few years, the discovery of new HPVs has accelerated[409,629].

NGS showed to be effective also in detecting low-copy HPVs and to discriminate multiple infections[400,630–632].

In this work, specific or degenerate primers targeting the L1 region of a broad spectrum of HPVs were used in combination with NGS, to discover new HPV types, particularly from the genus beta. This strategy implies the selective enrichment of PV sequences before the NGS, and similar approaches have been described in other works[209,400,401]. Around two-thirds of the reads generated, were related to PV sequences.

Different studies showed that beta HPV types might be involved in pre-malignant and malignant skin lesions, generating a growing interest in the biology of this HPV genus[183,633]. Species beta-3 HPV types have been identified in the skin, and mucosal epithelia[186,401] and functional studies have highlighted some biological similarities between beta-3 and mucosal HR HPV types. Beta-3 HPV49 showed transforming activity in primary human keratinocytes, and shares some features with HPV16[189,190].

Therefore, one major objective of our study was to expand the species beta-3, which includes only 4 HPV types, by using beta-3 consensus and degenerate primers.

With our protocol, that combines the use of different PCR protocols for the detection of HPV sequences and NGS, a total of 105 putative new PVs were discovered, in addition to the detection of 296 known PV types. A substantial number of beta and gamma HPV types were identified in the oral cavity, supporting the hypothesis of a possible mucosal tropism. Nevertheless, environmental contamination of the oral cavity cannot be excluded. Also, numerous other sequences related to unclassified and non-human PVs were detected in skin and oral samples. The presence of non-human PVs in the skin and oral samples can be due to environmental contamination. Though, cross-species transmission of PVs between animals and humans may also be hypothesized[634,635], although PVs are usually considered to be highly host-restricted (with a few exceptions). Interspecies transmission events have been considered, for example, in studies where sequences related to bovine PVs have been found in horses and other equids[636,637]. Different cases of cross-species transmission of PVs have been reported. For example, between bat species[638], between rhesus and cynomolgus macaques[639] and, between humans and cats[640,641]; however, additional studies are needed to confirm the latter. Moreover, the definition of “non-human” PV genera needs to be interpreted with attention as they may also include some HPVs. Also, alpha and beta genera include a few non-human primate PVs[168,642]. The results described in this study are based on the identification of the sequences using the RAxML-EPA classification. A total of 105 putative new PVs (including 29 beta, 32 gamma, two alpha, and one mu PV types) were identified in this study. An additional 24 PVs not belonging to any of the five PV genera that contain HPVs were detected. Among the 105 putative new PVs, 17 (16.2%) were assigned to taxonomically un-classified PVs and thus may be representative of putative new genera. The taxonomic identification performed in this study must be considered carefully since only small portions of putative new PV genomes have been obtained. Moreover, the results of the Blastn analysis refer only to the portion of the sequence that is aligned by the algorithm. Also, the

MegaBlast results used to define the known, and putative new sequences must be interpreted carefully since the definition of a novel PV type is based on the full L1 ORF length.

In this work, different PCR protocols were used to amplify PV sequences in two different human samples, with different efficacies in detecting putative new PVs and known PVs. Four new beta-3-related sequences were identified in skin samples using the beta-3-1 and beta-3-2 protocols, (according to the RAxML-EPA classification), potentially expanding the beta-3 group to 8 PV types. In vitro experiments are required to characterize these PV types, to understand whether these types share biological features with HPV49[189,190]. A broad range of PV types was detected in skin and oral samples using the CUT primers, including alpha PV types, as previously reported[383]. Instead, the original FAP protocol was much less effective in identifying alpha types. In oral samples, using the FAPM1 and FAPM2 protocols, the largest number of putative new PVs was identified.

In contrast, in skin samples, the largest number of putative new PVs was identified using the CUT primers. In skin samples, the FAPM1 and CUT protocols showed a good capability in detecting new PV types, belonging to non-human PV genera. In this study, a total of 62 putative new beta and gamma HPV types were detected, as well as 24 putative non-human PVs, in both skin and oral samples. The role of HPVs belonging to the gamma genus in human diseases is not clear. Nevertheless, HPV197, from species gamma-24, that constitute a small group of viruses has recently been found in human skin cancer samples[208,209]. The development of primers, based on the known gamma-24 types, can lead to the identification of new related PV types if any exist. Further experiments are required to evaluate the possible transforming activities of the new beta and gamma types identified in this study. In conclusion, this study describes a robust strategy for the detection of putative new PV types, using specific or degenerate primers and NGS.

The isolation and full characterization of a novel HPV type (i.e., HPV ICB1)[624], obtained in a preliminary study, confirms the effectiveness of our protocol as a first step for the isolation and full characterization of the novel HPVs.

The identification of novel HPV types remains of paramount importance to understand the role of HPVs in human diseases.

2. PVAmpliconFinder: a new workflow for the identification of PV sequences

PVAmpliconFinder, a novel workflow for the identification of known and putative new PV sequences, was developed. This workflow was specifically designed to analyze targeted sequencing data obtained through NGS. The input required by this workflow is the FASTQ file generated by the sequencing run. Different tabular and graphical output files are produced, describing the nature and abundance of PV-related sequences present in a complex mixture of host, phage, bacterial, and viral DNA. The sequences representative of known and putative new HPV types are identified and divided into different output files. Sequencing metrics and sequence details are provided, enabling the design of subsequent laboratory experiments for confirming the *in silico* findings. Unlike read-subtraction methods, PVAmpliconFinder aligns the sequences against the entire NCBI database. Thereby, removing host sequences may remove potentially new viral sequences that present some similarity to the host, and this has to be taken into account. Moreover, the use of degenerate primers helps in the amplification of a broad spectrum of viral sequences, and this can facilitate the discovery of novel HPV types. This workflow was specifically designed for NGS data, and different steps are performed to obtain higher quality sequences that are subsequently identified. The number of dereplicated sequences corresponding to each template is used to calculate an unnormalized abundance. A step of sequence clustering is performed to correct potential sequencing errors. The tool uses a 98% identity threshold for clustering, and this because 2% of dissimilarity from any known L1 gene is enough to define a new PV variant[168]. This threshold is a good compromise when the aim is to identify putative new PV types (with more than 10% of dissimilarity to known PVs in the L1 ORF).

The whole NCBI nt database is used for the MegaBlast analysis performed by PVAmpliconFinder. This step allows the retaining of only PV related sequences favoring the generation of unbiased results. A *de novo* assembly step is performed, after the identification of known and putative new HPV-related sequence, to reconstruct the longest sequence for each potential PV identified. PVAmpliconFinder uses sequence similarity and homology for an advanced identification and taxonomic classification of the sequences. The PaVe database, which also includes many "not referenced" genomes, is used with the Blastn algorithm to compute the sequence similarity[172]. Using MegaBlast and BlastN classifications, PVAmpliconFinder identified most of the PV sequences, but using this approach, a consistent number of sequences remain unclassified because they match against incomplete L1 cds.

Furthermore, pairwise alignment with a low percentage of similarity raises a concern about the pertinence of the results produced. For example, regarding the putative new sequences identified in the application example reported here, all having at least 15% of dissimilarity against their best match. Thus, RaxML-EPA[601] was used to obtain an alternative identification of the sequences. A multiple sequence alignment is used to infer evolutionary time and to generate a phylogenetic reference tree of selected species. After, the Parsimony-based Phylogeny-Aware Read alignment (PaPaRa) algorithm is used to find the best position of the different sequences into the reference multiple sequence alignment[605]. Thus, RaxML-EPA is used to find the best position of those sequences in the reference tree. Also, the evolutionary-based method used here to identify the sequence has his limitations, and the classification of the sequences has to be considered carefully. RaxML suffers from long-branch attraction error, where distant lineages are inferred to be close relatives because both have undergone a large number of changes. In our experiment, the classification by EPA, identified Erethizon dorsatum sequences, and this could be due to the limitations of this approach in identifying some sequences correctly. Erethizon dorsatum PV is a recently referenced but unclassified virus[643], presenting substantial differences from other known PVs on its L1 gene, and thought to represent a new genus in the Papillomaviridae family. Further

analyses are required to characterize the sequences and to verify their real correlation with Erethizon dorsatum PV.

Another important point to consider is the possibility of cross-contamination events between samples during library preparation, amplification, and sequencing, or environmental contamination that are hard to detect using *in silico* methods. For example, low-abundance sequences may also come from cross-contamination events. In general low-abundance sequences should be considered with caution. Environmental contamination may explain the presence of non-human PV in human samples. Nevertheless, cross-contamination between species has already been described[634,635] and thus cannot be excluded. Few bioinformatics methods are available for the identification of HPV sequences in NGS data, and they are often restricted to a panel of already well-characterized PV types[644]. The use of PVAmpliconFinder, in combination with primers specifically developed for the amplification of PV sequences, and NGS, can represent a valid approach for the identification of novel PV types. The outputs generated by PVAmpliconFinder provide the necessary information to select the most promising putative new PV sequences that may be validated by further wet-lab approaches. Additionally, PVAmpliconFinder can be easily modified and applied to other viral families. PVAmpliconFinder could also be used in clinical research settings.

Specific aim #2

1. Full genomic characterization of a novel Beta-2 Human Papillomavirus

In specific aim #1 of the present work, putative new PVs were identified using different PCR protocols, designed for the amplification of HPV sequences, in combination with NGS.

The NGS analysis produced partial L1 ORF sequences representative of 105 putative novel PVs. The complete sequence of the L1 ORF is required to define a putative novel HPV type as truly new[168]. Therefore, to validate the effectiveness of the strategy used to detect novel HPV types, starting from the partial L1 ORF sequence of HPV ICB2 (99 bp), the putative novel HPV type obtained in specific aim #1 of the present work, the whole genome of this virus was reconstructed.

To obtain the complete viral genome, first, long-range PCR was performed after RCA, on the original skin sample, positive for HPV ICB2.

The resulting amplicon of approximately 8 kb was then cloned, and the sequence of the whole genome was obtained by Sanger sequencing using a primer-walking strategy.

The viral genome was covered at least twice, to identify and correct sequencing errors. Thirty-one sequences were generated and aligned to reconstruct the whole genome using CAP3 *de novo* assembling tool.

The clone has been submitted to the International Human Papillomavirus Reference Center in Stockholm (www.hpvcenter.se), and the virus name HPV227 was assigned.

The L1 open reading frame (ORF) of HPV227 showed 87.9% nucleotide identity with its closest relative, HPV37, from species beta-2 of the genus betapapillomavirus. HPV227 thus constitutes a novel human betapapillomavirus, sharing less than 90% nucleotide sequence identity with the closest HPV type in the L1 ORF[170].

This result confirms the effectiveness of our strategy in detecting novel HPV types and contributes to the expansion of our knowledge about the diversity of the genus betapapillomavirus.

Specific aim #3

1. MinION sequencing for the reconstruction of HPV genomes

In specific aim #1 of the present work, the partial L1 ORF sequence of HPV ICB2, a putative novel HPV type, was discovered using broad-spectrum primers and NGS. In specific aim #2, the whole genome of HPV ICB2, identified as a novel beta-2 HPV type, was obtained using a primer-walking strategy and Sanger sequencing and the official name “HPV227” was assigned to this virus.

The strategies used to identify and reconstruct the entire genome of HPV227, based on Sanger sequencing and NGS, have been described previously[645–651]. These methodologies are low-throughput and time-consuming, respectively. Moreover, whereas such technologies generate short reads of only up to a few hundred nucleotides, the use of metagenomics data for viral genome reconstruction can lead to the artefactual assembly of chimeric genomes. This risk is particularly high when several related HPVs are present in the sample[652–656]. Thus, the last aim of this study was to identify a novel strategy for the rapid and effective characterization of novel HPV types.

For this purpose, we evaluated the capability of MinION nanopore sequencer, developed by Oxford Nanopore Technologies (ONT), to sequence the whole genome of HPV227.

Many recent studies have reported the use of this technology for the sequencing of viral genomes from hepatitis B[652], simian immunodeficiency virus[657], Ebola virus[658], hepatitis C[659], herpesvirus[660], poxvirus[661], parapoxvirus[662], or Zika virus[663], with accuracies of up to 99%. A novel species of papillomavirus in giraffe lesions was also identified using this technology[664]. Moreover, several studies have used long-reads sequencing for transcriptomic studies of herpesvirus[665–668], poxvirus[669], and baculovirus[670,671], or viral genome detection or analysis of avipoxvirus[672], henipavirus[673], and poxvirus[674].

We established two sequencing protocols (A and B), both of which based on the sequencing of three barcoded libraries, either independently or pooled, to obtain an accurate sequence of the whole HPV227 genome and fully explore the potential of the MinION technology.

A novel bioinformatics pipeline was developed, allowing the reconstruction of the whole genome of HPV227 from both protocols with error rates of only 0.07 and 0.11%, respectively. DraftPolisher, an in-house script that works at the consensus level, and Nanopolish, a recommended and widely used tool, performing a signal-level polishing[652,654,675–677], were used to generate high-quality consensus sequences.

Among the five errors (5 nucleotide gaps) in protocol A and the eight errors (8 nucleotide gaps) in protocol B, five (nucleotide positions 882, 2165, 2612, 3491, and 4182) corresponded to positions with the lowest percentage of similarities to HPV227 (Figure 49 B), thus representing potential systematic errors. Furthermore, sequencing errors may also be introduced during the bioinformatics analysis (*e.g.*, base-calling or polishing steps), as reported elsewhere[652]. Though, the use of improved sequencing flow cells and chemistries and the development of ONT technologies with much better accuracy may help in generating higher quality raw data, reducing the error rate[627]. All the nucleotide gaps (*N*=13) identified in this study were due to the insertion or deletion of a single nucleotide within homopolymeric A (*N*=2), T (*N*=2), G (*N*=5), and C (*N*=4) stretches. This base-calling limitation of the Oxford Nanopore technology has already been reported [627]. The potential introduction of errors occurred during the PCR steps, was verified, and no artefacts were found to be attributable to those amplification steps.

The results of this study also showed the importance of using filtering and assembly steps to avoid misclassification of the sequences. Before performing these steps, up to 21.3% of the reads were potentially misclassified as belonging to archaea, eukaryotes, or bacteria, while BLAST analysis, performed after upstream filtering and assembly steps, showed that at least 99.0% of the sequences were representative of HPV227.

In protocol B, only 560 nanopores were available for the sequencing run. Despite this low number of active nanopores, the reconstruction of the whole genome of HPV227 with a per cent identity of 99.89% was obtained, with 83% of the passed reads generated within the first 12 hours of the sequencing run.

According to our results, we can hypothesize that the use of a flow cell with at least 800 active nanopores and a single run of up to 48 hours, should allow the generation of good-quality sequencing data for the reconstruction of up to four viral genomes in triplicate. Furthermore, our data showed that a sequencing run of 3 hours was determined as the shortest run time sufficient to assemble the full-length viral genome from both protocols, with percentages of identity to HPV227 exceeding 99.9%.

The protocol described in this study can be performed in less than two weeks, representing an alternative to time-consuming methodologies generally used for the characterization of HPV genomes. The RCA step requires one day, followed by another day for the long-range PCR using HPV227-specific outward-directed primers, agarose gel purification, and library preparation. The total duration of the protocol depends on the sequencing run time and on the bioinformatics analysis, which can last several days (up to 8 days). In our analysis, SPAdes required several days to perform the *de novo* assembling, thus constituting the longer part of our protocol.

Between each of the runs, a washing step was performed to remove traces of DNA that could contaminate the following run. However, traces of DNA from the previous runs was detected in run 2, 3, and 4 of our study, and this phenomenon was also described in previous publications[626,627].

However, data generated from runs 2 and 3 (Protocol A) showed that only 9.2% and 4.9% of the barcoded reads, respectively, were identified as contaminants from previous runs. Thus the impact of this technical limitation on our study is minimal. The use of different barcoded libraries

on the same flow cells when applying this protocol is thus warranted. Moreover, our data confirmed that launching consecutive runs using the same flow cell leads to a drastic decrease of active nanopores. Since too many nanopores become inactive after consecutive runs, a single multiplexed run with barcoded libraries might be considered to obtain reads of sufficient quality and quantity to reconstruct multiple PV genomes.

An HPV is defined as a new type when its L1 ORF shows at least 10% of dissimilarity to the L1 ORF of the closest known type. Therefore, an error rate of around 0.1% using the MinION sequencing is tolerable to identify known or new HPV types. Additionally, this technology can be easily integrated into our protocol for the isolation and full characterization of novel HPV genomes[647,678,679].

In conclusion, with this work, we showed that the MinION nanopore sequencer is an effective and rapid strategy for long-reads sequencing, allowing the full characterization of novel HPV genomes.

Moreover, we proposed a robust strategy to analyze the MinION sequencing data, obtaining the whole genome of HPV227 with a very low error rate. We believe that future improvements in this technology and specific data analysis tools will allow the full and rapid characterization of new HPVs.

Conclusions

In specific aim #1 of the present study, we used broad-spectrum PCR primers to amplify a portion of the HPV L1 ORF, combined with next-generation sequencing, for the identification of putative new HPVs. This strategy led to the discovery of 105 putative new PV types in addition to 296 known types, thus providing useful information about the viral distribution in the oral cavity and skin.

In the second phase of our work, the pipeline used to analyze the NGS sequencing data was further improved and optimized to identify putative new HPV types.

In specific aim #2 of the present work, starting from a partial L1 region sequence obtained in the NGS analysis (*i.e.,* 99 bp fragment representative of the putative novel HPV-ICB2), using long-range PCR with specific outward-directed primers and Sanger sequencing, the full-length genome of the novel papillomavirus type was obtained. The official name “HPV227” was assigned to this new beta-2 papillomavirus.

Finally, in specific aim #3 of the present work, we used the MinION technology to sequence the whole genome of HPV227, starting from the original skin sample where this virus was discovered. We specifically developed a bioinformatics analysis for *de novo* reconstruction of HPV genomes from MinION sequencing data. Using this strategy, the whole genome of HPV227 was obtained, with a pairwise identity to the reference genome of the virus, exceeding 99%.

Overall, this thesis describes a reliable protocol for the identification and full characterization of novel HPV genomes. Moreover, the discovery of 105 putative novel PV types expands our knowledge of this family of viruses. Further analyses will be required for a complete characterization of all these new viruses and to investigate their potential role in human diseases.

Bibliography

1. Lecoq, H. [Discovery of the first virus, the tobacco mosaic virus: 1892 or 1898?]. *C. R. Acad. Sci. III, Sci. Vie* **324**, 929–933 (2001).
2. Cai, Q. & Yuan, Z. Overview of Infectious Causes of Human Cancers. *Adv. Exp. Med. Biol.* **1018**, 1–9 (2017).
3. Javier, R. T. & Butel, J. S. The history of tumor virology. *Cancer Res.* **68**, 7693–7706 (2008).
4. Rous, P. A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. *J. Exp. Med.* **13**, 397–411 (1911).
5. Rous, P. A transmissible avian neoplasm (sarcoma of the common fowl). 1910. *Clin. Orthop. Relat. Res.* 3–8 (1993).
6. Epstein, M. A., Achong, B. G. & Barr, Y. M. VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *Lancet* **1**, 702–703 (1964).
7. Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311–1315 (1984).
8. Warren, J. R. & Marshall, B. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet* **1**, 1273–1275 (1983).
9. zur Hausen, H., Meinhof, W., Scheiber, W. & Bornkamm, G. W. Attempts to detect virus-secific DNA in human tumors. I. Nucleic acid hybridizations with complementary RNA of human wart virus. *Int. J. Cancer* **13**, 650–656 (1974).
10. Feng, H., Shuda, M., Chang, Y. & Moore, P. S. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* **319**, 1096–1100 (2008).
11. Shuda, M. *et al.* T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 16272–16277 (2008).
12. Chang, Y. *et al.* Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865–1869 (1994).

13. de Martel, C. *et al.* Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.* **13**, 607–615 (2012).
14. Plummer, M. *et al.* Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob Health* **4**, e609-616 (2016).
15. Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* **118**, 3030–3044 (2006).
16. Pisani, P., Parkin, D. M., Muñoz, N. & Ferlay, J. Cancer and infection: estimates of the attributable fraction in 1990. *Cancer Epidemiol. Biomarkers Prev.* **6**, 387–400 (1997).
17. de Martel, C., Georges, D., Bray, F., Ferlay, J. & Clifford, G. M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* **8**, e180–e190 (2020).
18. Plummer, M. *et al.* Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob Health* **4**, e609-616 (2016).
19. zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* **2**, 342–350 (2002).
20. McGlynn, K. A. & London, W. T. Epidemiology and natural history of hepatocellular carcinoma. *Best Pract Res Clin Gastroenterol* **19**, 3–23 (2005).
21. Andersson, J. An Overview of Epstein-Barr Virus: from Discovery to Future Directions for Treatment and Prevention. *Herpes* **7**, 76–82 (2000).
22. Li, S., Bai, L., Dong, J., Sun, R. & Lan, K. Kaposi's Sarcoma-Associated Herpesvirus: Epidemiology and Molecular Biology. *Adv. Exp. Med. Biol.* **1018**, 91–127 (2017).
23. Chan, C.-P., Kok, K.-H. & Jin, D.-Y. Human T-Cell Leukemia Virus Type 1 Infection and Adult T-Cell Leukemia. *Adv. Exp. Med. Biol.* **1018**, 147–166 (2017).
24. MacDonald, M. & You, J. Merkel Cell Polyomavirus: A New DNA Virus Associated with Human Cancer. *Adv. Exp. Med. Biol.* **1018**, 35–56 (2017).

25. Ahmed, N. & Sechi, L. A. Helicobacter pylori and gastroduodenal pathology: new threats of the old friend. *Ann. Clin. Microbiol. Antimicrob.* **4**, 1 (2005).
26. Mostafa, M. H., Sheweita, S. A. & O'Connor, P. J. Relationship between Schistosomiasis and Bladder Cancer. *Clin Microbiol Rev* **12**, 97–111 (1999).
27. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359-386 (2015).
28. Pisani, P., Parkin, D. M., Muñoz, N. & Ferlay, J. Cancer and infection: estimates of the attributable fraction in 1990. *Cancer Epidemiol. Biomarkers Prev.* **6**, 387–400 (1997).
29. Harper, D. M. & DeMars, L. R. HPV vaccines - A review of the first decade. *Gynecol. Oncol.* **146**, 196–204 (2017).
30. Bray, F. *et al.* Cancer Incidence in Five Continents: Inclusion criteria, highlights from Volume X and the global status of cancer registration. *Int. J. Cancer* **137**, 2060–2071 (2015).
31. Chaturvedi, A. K. *et al.* Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States. *JCO* **29**, 4294–4301 (2011).
32. Gillison, M. L., Chaturvedi, A. K. & Lowy, D. R. HPV prophylactic vaccines and the potential prevention of noncervical cancers in both men and women. *Cancer* **113**, 3036–3046 (2008).
33. von Krogh, G. Management of anogenital warts (condylomata acuminata). *Eur J Dermatol* **11**, 598–603; quiz 604 (2001).
34. von Krogh, G. *et al.* European guideline for the management of anogenital warts. *Int J STD AIDS* **12 Suppl 3**, 40–47 (2001).
35. Armstrong, L. R. *et al.* Incidence and prevalence of recurrent respiratory papillomatosis among children in Atlanta and Seattle. *Clin. Infect. Dis.* **31**, 107–109 (2000).
36. Joura, E. A. *et al.* A 9-valent HPV vaccine against infection and intraepithelial neoplasia in women. *N. Engl. J. Med.* **372**, 711–723 (2015).
37. Monie, A., Hung, C.-F., Roden, R. & Wu, T.-C. Cervarix: a vaccine for the prevention of HPV 16, 18-associated cervical cancer. *Biologics* **2**, 97–105 (2008).

38. World Health Organization. Electronic address: sageexecsec@who.int. Human papillomavirus vaccines: WHO position paper, May 2017-Recommendations. *Vaccine* **35**, 5753–5755 (2017).

39. L.M. Fernandez, D.M. Harper, M.E. Pendleton & R.B. Wright,. Chapter 29 - bivalent HPV vaccine approved for cervical cancer prevention in females. in *A. Ayhan, N. Reed, M. Gultekin, P. Dursun (Eds.), Textbook of Gynaecological Oncology, third ed.Gunes Publishing, Ankara 2016, pp. 247–278*.

40. A.S. LaJoie, L.M. Fernandez, M.E. Pendleton & D.M. Harper. Chapter 30 - quadrivalent and nonavalent HPV vaccine approved for males and females for HPV associated diseases. in *A. Ayhan, N. Reed, M. Gultekin, P. Dursun (Eds.), Textbook of Gynaecological Oncology, third ed.Gunes Publishing, Ankara 2016, pp. 279–308.*

41. Meggiolaro, A. *et al.* The role of Pap test screening against cervical cancer: a systematic review and meta-analysis. *Clin Ter* **167**, 124–139 (2016).

42. Horn, J. *et al.* Reduction of cervical cancer incidence within a primary HPV screening pilot project (WOLPHSCREEN) in Wolfsburg, Germany. *Br. J. Cancer* **120**, 1015–1022 (2019).

43. Goodman, A. HPV testing as a screen for cervical cancer. *BMJ* **350**, h2372 (2015).

44. Bonanni, P., Boccalini, S. & Bechini, A. Efficacy, duration of immunity and cross protection after HPV vaccination: a review of the evidence. *Vaccine* **27 Suppl 1**, A46-53 (2009).

45. Hanson, C. M., Eckert, L., Bloem, P. & Cernuschi, T. Gavi HPV Programs: Application to Implementation. *Vaccines (Basel)* **3**, 408–419 (2015).

46. Hall, M. T. *et al.* The projected timeframe until cervical cancer elimination in Australia: a modelling study. *Lancet Public Health* **4**, e19–e27 (2019).

47. Ng, J. *et al.* Human Papillomavirus Vaccination Coverage Among Female Adolescents in Managed Care Plans - United States, 2013. *MMWR Morb. Mortal. Wkly. Rep.* **64**, 1185–1189 (2015).

48. Centers for Disease Control and Prevention (CDC). Recommendations on the use of quadrivalent human papillomavirus vaccine in males--Advisory Committee on Immunization Practices (ACIP), 2011. *MMWR Morb. Mortal. Wkly. Rep.* **60**, 1705–1708 (2011).

49. Fagot, J.-P., Boutrelle, A., Ricordeau, P., Weill, A. & Allemand, H. HPV vaccination in France: uptake, costs and issues for the National Health Insurance. *Vaccine* **29**, 3610–3616 (2011).

50. Soe, N. N. *et al.* Should human papillomavirus vaccination target women over age 26, heterosexual men and men who have sex with men? A targeted literature review of cost-effectiveness. *Hum Vaccin Immunother* **14**, 3010–3018 (2018).

51. Italian Ministry of Health. Vaccinazione contro il papilloma virus (HPV) - Coperture vaccinali. (2017).

52. Crosbie, E. J., Einstein, M. H., Franceschi, S. & Kitchener, H. C. Human papillomavirus and cervical cancer. *Lancet* **382**, 889–899 (2013).

53. Harden, M. E. & Munger, K. Human papillomavirus molecular biology. *Mutat Res Rev Mutat Res* **772**, 3–12 (2017).

54. Tommasino, M. The human papillomavirus family and its role in carcinogenesis. *Semin. Cancer Biol.* **26**, 13–21 (2014).

55. Gheit, T. Mucosal and Cutaneous Human Papillomavirus Infections and Cancer Biology. *Front Oncol* **9**, 355 (2019).

56. Guan, J. *et al.* Cryoelectron Microscopy Maps of Human Papillomavirus 16 Reveal L2 Densities and Heparin Binding Site. *Structure* **25**, 253–263 (2017).

57. Doorbar, J., Egawa, N., Griffin, H., Kranjec, C. & Murakami, I. Human papillomavirus molecular biology and disease association. *Rev. Med. Virol.* **25 Suppl 1**, 2–23 (2015).

58. Knipe, DM., Howley, PM. *Fields virology*. (Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins Health, 2013).

59. Van Doorslaer, K. & McBride, A. A. Molecular archeological evidence in support of the repeated loss of a papillomavirus gene. *Sci Rep* **6**, 33028 (2016).

60. Chen, Z., Schiffman, M., Herrero, R., Desalle, R. & Burk, R. D. Human papillomavirus (HPV) types 101 and 103 isolated from cervicovaginal cells lack an E6 open reading frame (ORF) and are related to gamma-papillomaviruses. *Virology* **360**, 447–453 (2007).

61. Nobre, R. J. *et al.* E7 oncoprotein of novel human papillomavirus type 108 lacking the E6 gene induces dysplasia in organotypic keratinocyte cultures. *J. Virol.* **83**, 2907–2916 (2009).

62. McBride, A. A. The papillomavirus E2 proteins. *Virology* **445**, 57–79 (2013).

63. Schmitt, A. *et al.* The primary target cells of the high-risk cottontail rabbit papillomavirus colocalize with hair follicle stem cells. *J. Virol.* **70**, 1912–1922 (1996).

64. Kines, R. C., Thompson, C. D., Lowy, D. R., Schiller, J. T. & Day, P. M. The initial steps leading to papillomavirus infection occur on the basement membrane prior to cell surface binding. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 20458–20463 (2009).

65. Schiller, J. T., Day, P. M. & Kines, R. C. Current understanding of the mechanism of HPV infection. *Gynecol. Oncol.* **118**, S12-17 (2010).

66. Herfs, M. *et al.* A discrete population of squamocolumnar junction cells implicated in the pathogenesis of cervical cancer. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10516–10521 (2012).

67. Kurita, T. Normal and Abnormal Epithelial Differentiation in the Female Reproductive Tract. *Differentiation* **82**, 117–126 (2011).

68. Fluhmann, C. F. Comparative studies of squamous metaplasia of the cervix uteri and endometrium. *Am. J. Obstet. Gynecol.* **68**, 1447–1463 (1954).

69. Beckmann, Charles R B A, Laube, Douglas, Herbert, William, Ling, Frank & Smith, Roger. *Obstetrics and Gynecology (7th ed.).* (2013).

70. Castle, P. E. *et al.* A prospective study of age trends in cervical human papillomavirus acquisition and persistence in Guanacaste, Costa Rica. *J. Infect. Dis.* **191**, 1808–1816 (2005).

71. Moscicki, A.-B., Ellenberg, J. H., Farhat, S. & Xu, J. Persistence of human papillomavirus infection in HIV-infected and -uninfected adolescent girls: risk factors and differences, by phylogenetic type. *J. Infect. Dis.* **190**, 37–45 (2004).

72. Rautava, J. & Syrjänen, S. Biology of human papillomavirus infections in head and neck carcinogenesis. *Head Neck Pathol* **6 Suppl 1**, S3-15 (2012).

73. Broniarczyk, J., Ring, N., Massimi, P., Giacca, M. & Banks, L. HPV-16 virions can remain infectious for 2 weeks on senescent cells but require cell cycle re-activation to allow virus entry. *Sci Rep* **8**, 811 (2018).

74. Cerqueira, C., Samperio Ventayol, P., Vogeley, C. & Schelhaas, M. Kallikrein-8 Proteolytically Processes Human Papillomaviruses in the Extracellular Space To Facilitate Entry into Host Cells. *J. Virol.* **89**, 7038–7052 (2015).

75. Culp, T. D., Budgeon, L. R., Marinkovich, M. P., Meneguzzi, G. & Christensen, N. D. Keratinocyte-secreted laminin 5 can function as a transient receptor for human papillomaviruses by binding virions and transferring them to adjacent cells. *J. Virol.* **80**, 8940–8950 (2006).

76. Bronnimann, M. P. *et al.* Furin Cleavage of L2 during Papillomavirus Infection: Minimal Dependence on Cyclophilins. *J. Virol.* **90**, 6224–6234 (2016).

77. Rubio, I. *et al.* The N-terminal region of the human papillomavirus L2 protein contains overlapping binding sites for neutralizing, cross-neutralizing and non-neutralizing antibodies. *Virology* **409**, 348–359 (2011).

78. Schelhaas, M. *et al.* Human papillomavirus type 16 entry: retrograde cell surface transport along actin-rich protrusions. *PLoS Pathog.* **4**, e1000148 (2008).

79. Schelhaas, M. *et al.* Entry of human papillomavirus type 16 by actin-dependent, clathrin- and lipid raft-independent endocytosis. *PLoS Pathog.* **8**, e1002657 (2012).

80. Siddiqa, A., Broniarczyk, J. & Banks, L. Papillomaviruses and Endocytic Trafficking. *Int J Mol Sci* **19**, (2018).

81. Bienkowska-Haba, M., Williams, C., Kim, S. M., Garcea, R. L. & Sapp, M. Cyclophilins facilitate dissociation of the human papillomavirus type 16 capsid protein L1 from the L2/DNA complex following virus entry. *J. Virol.* **86**, 9875–9887 (2012).

82. DiGiuseppe, S., Bienkowska-Haba, M., Guion, L. G. M., Keiffer, T. R. & Sapp, M. Human Papillomavirus Major Capsid Protein L1 Remains Associated with the Incoming Viral Genome throughout the Entry Process. *J. Virol.* **91**, (2017).

83. Popa, A. *et al.* Direct binding of retromer to human papillomavirus type 16 minor capsid protein L2 mediates endosome exit during viral infection. *PLoS Pathog.* **11**, e1004699 (2015).

84. McNally, K. E. *et al.* Retriever is a multiprotein complex for retromer-independent endosomal cargo recycling. *Nat. Cell Biol.* **19**, 1214–1225 (2017).

85. Bergant Marušič, M., Ozbun, M. A., Campos, S. K., Myers, M. P. & Banks, L. Human papillomavirus L2 facilitates viral escape from late endosomes via sorting nexin 17. *Traffic* **13**, 455–467 (2012).

86. Bergant, M., Peternel, Š., Pim, D., Broniarczyk, J. & Banks, L. Characterizing the spatio-temporal role of sorting nexin 17 in human papillomavirus trafficking. *J. Gen. Virol.* **98**, 715–725 (2017).

87. Pim, D., Broniarczyk, J., Bergant, M., Playford, M. P. & Banks, L. A Novel PDZ Domain Interaction Mediates the Binding between Human Papillomavirus 16 L2 and Sorting Nexin 27 and Modulates Virion Trafficking. *J. Virol.* **89**, 10145–10155 (2015).

88. DiGiuseppe, S., Bienkowska-Haba, M. & Sapp, M. Human Papillomavirus Entry: Hiding in a Bubble. *J. Virol.* **90**, 8032–8035 (2016).

89. Pyeon, D., Pearce, S. M., Lank, S. M., Ahlquist, P. & Lambert, P. F. Establishment of human papillomavirus infection requires cell cycle progression. *PLoS Pathog.* **5**, e1000318 (2009).

90. Aydin, I. *et al.* Large scale RNAi reveals the requirement of nuclear envelope breakdown for nuclear import of human papillomaviruses. *PLoS Pathog.* **10**, e1004162 (2014).

91. Ascoli, C. A. & Maul, G. G. Identification of a novel nuclear domain. *J. Cell Biol.* **112**, 785–795 (1991).

92. Day, P. M., Baker, C. C., Lowy, D. R. & Schiller, J. T. Establishment of papillomavirus infection is enhanced by promyelocytic leukemia protein (PML) expression. *Proc Natl Acad Sci U S A* **101**, 14252–14257 (2004).

93. Tavalai, N., Adler, M., Scherer, M., Riedl, Y. & Stamminger, T. Evidence for a dual antiviral role of the major nuclear domain 10 component Sp100 during the immediate-early and late phases of the human cytomegalovirus replication cycle. *J. Virol.* **85**, 9447–9458 (2011).

94. Stepp, W. H., Meyers, J. M. & McBride, A. A. Sp100 provides intrinsic immunity against human papillomavirus infection. *mBio* **4**, e00845-00813 (2013).

95. Florin, L., Schäfer, F., Sotlar, K., Streeck, R. E. & Sapp, M. Reorganization of nuclear domain 10 induced by papillomavirus capsid protein l2. *Virology* **295**, 97–107 (2002).

96. Kivipõld, P., Võsa, L., Ustav, M. & Kurg, R. DAXX modulates human papillomavirus early gene expression and genome replication in U2OS cells. *Virol. J.* **12**, 104 (2015).

97. Doorbar, J. *et al.* The biology and life-cycle of human papillomaviruses. *Vaccine* **30 Suppl 5**, F55-70 (2012).

98. McKinney, C. C., Kim, M. J., Chen, D. & McBride, A. A. Brd4 Activates Early Viral Transcription upon Human Papillomavirus 18 Infection of Primary Keratinocytes. *mBio* **7**, (2016).

99. Sanders, C. M. & Stenlund, A. Recruitment and loading of the E1 initiator protein: an ATP-dependent process catalysed by a transcription factor. *EMBO J.* **17**, 7044–7055 (1998).

100. Gauson, E. J. *et al.* Evidence supporting a role for TopBP1 and Brd4 in the initiation but not continuation of human papillomavirus 16 E1/E2-mediated DNA replication. *J. Virol.* **89**, 4980–4991 (2015).

101. Dreer, M. *et al.* Interaction of NCOR/SMRT Repressor Complexes with Papillomavirus E8^E2C Proteins Inhibits Viral Replication. *PLoS Pathog.* **12**, e1005556 (2016).

102. Dreer, M., van de Poel, S. & Stubenrauch, F. Control of viral replication and transcription by the papillomavirus E8^E2 protein. *Virus Res.* **231**, 96–102 (2017).

103. McPhillips, M. G., Oliveira, J. G., Spindler, J. E., Mitra, R. & McBride, A. A. Brd4 is required for e2-mediated transcriptional activation but not genome partitioning of all papillomaviruses. *J. Virol.* **80**, 9530–9543 (2006).

104. Bastien, N. & McBride, A. A. Interaction of the papillomavirus E2 protein with mitotic chromosomes. *Virology* **270**, 124–134 (2000).

105. Jang, M. K., Shen, K. & McBride, A. A. Papillomavirus genomes associate with BRD4 to replicate at fragile sites in the host genome. *PLoS Pathog.* **10**, e1004117 (2014).

106. McBride, A. A. & Jang, M. K. Current understanding of the role of the Brd4 protein in the papillomavirus lifecycle. *Viruses* **5**, 1374–1394 (2013).

107. Abroi, A., Ilves, I., Kivi, S. & Ustav, M. Analysis of chromatin attachment and partitioning functions of bovine papillomavirus type 1 E2 protein. *J. Virol.* **78**, 2100–2113 (2004).

108. Rogers, A., Waltke, M. & Angeletti, P. C. Evolutionary variation of papillomavirus E2 protein and E2 binding sites. *Virol. J.* **8**, 379 (2011).

109. Parish, J. L., Bean, A. M., Park, R. B. & Androphy, E. J. ChlR1 is required for loading papillomavirus E2 onto mitotic chromosomes and viral genome maintenance. *Mol. Cell* **24**, 867–876 (2006).

110. Parish, J. L. *et al.* The DNA helicase ChlR1 is required for sister chromatid cohesion in mammalian cells. *J. Cell. Sci.* **119**, 4857–4865 (2006).

111. Harris, L., McFarlane-Majeed, L., Campos-León, K., Roberts, S. & Parish, J. L. The Cellular DNA Helicase ChlR1 Regulates Chromatin and Nuclear Matrix Attachment of the Human Papillomavirus 16 E2 Protein and High-Copy-Number Viral Genome Establishment. *J. Virol.* **91**, (2017).

112. Bentley, P., Tan, M. J. A., McBride, A. A., White, E. A. & Howley, P. M. The SMC5/6 Complex Interacts with the Papillomavirus E2 Protein and Influences Maintenance of Viral Episomal DNA. *J. Virol.* **92**, (2018).

113. Thomas, J. T., Hubert, W. G., Ruesch, M. N. & Laimins, L. A. Human papillomavirus type 31 oncoproteins E6 and E7 are required for the maintenance of episomes during the viral life cycle in normal human keratinocytes. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8449–8454 (1999).

114. Lorenz, L. D., Rivera Cardona, J. & Lambert, P. F. Inactivation of p53 rescues the maintenance of high risk HPV DNA genomes deficient in expression of E6. *PLoS Pathog.* **9**, e1003717 (2013).

115. Moody, C. A. & Laimins, L. A. Human papillomavirus oncoproteins: pathways to transformation. *Nat. Rev. Cancer* **10**, 550–560 (2010).

116. Maglennon, G. A., McIntosh, P. & Doorbar, J. Persistence of viral DNA in the epithelial basal layer suggests a model for papillomavirus latency following immune regression. *Virology* **414**, 153–163 (2011).

117. Hummel, M., Hudson, J. B. & Laimins, L. A. Differentiation-induced and constitutive transcription of human papillomavirus type 31b in cell lines containing viral episomes. *J. Virol.* **66**, 6070–6080 (1992).

118. Fehrmann, F., Klumpp, D. J. & Laimins, L. A. Human papillomavirus type 31 E5 protein supports cell cycle progression and activates late viral functions upon epithelial differentiation. *J. Virol.* **77**, 2819–2831 (2003).

119. Longworth, M. S. & Laimins, L. A. Pathogenesis of human papillomaviruses in differentiating epithelia. *Microbiol. Mol. Biol. Rev.* **68**, 362–372 (2004).

120. Khan, J. *et al.* Role of calpain in the formation of human papillomavirus type 16 E1^E4 amyloid fibers and reorganization of the keratin network. *J. Virol.* **85**, 9984–9997 (2011).

121. Pentland, I. *et al.* Disruption of CTCF-YY1-dependent looping of the human papillomavirus genome activates differentiation-induced viral oncogene transcription. *PLoS Biol.* **16**, e2005752 (2018).

122. Butz, K. *et al.* siRNA targeting of the viral E6 oncogene efficiently kills human papillomavirus-positive cancer cells. *Oncogene* **22**, 5938–5945 (2003).

123. Fujii, T. *et al.* Intratumor injection of small interfering RNA-targeting human papillomavirus 18 E6 and E7 successfully inhibits the growth of cervical cancer. *Int. J. Oncol.* **29**, 541–548 (2006).

124. Bernard, B. A. *et al.* The human papillomavirus type 18 (HPV18) E2 gene product is a repressor of the HPV18 regulatory region in human keratinocytes. *J Virol* **63**, 4317–4324 (1989).

125. Thierry, F. & Yaniv, M. The BPV1-E2 trans-acting protein can be either an activator or a repressor of the HPV18 regulatory region. *EMBO J* **6**, 3391–3397 (1987).

126. Jeon, S. & Lambert, P. F. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical carcinogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1654–1658 (1995).

127. Akagi, K. *et al.* Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199 (2014).

128. Kadaja, M., Isok-Paas, H., Laos, T., Ustav, E. & Ustav, M. Mechanism of genomic instability in cells infected with the high-risk human papillomaviruses. *PLoS Pathog.* **5**, e1000397 (2009).

129. Kadaja, M. *et al.* Genomic instability of the host cell induced by the human papillomavirus replication machinery. *EMBO J.* **26**, 2180–2191 (2007).

130. McBride, A. A. & Warburton, A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog* **13**, (2017).

131. Vinokurova, S. *et al.* Type-dependent integration frequency of human papillomavirus genomes in cervical lesions. *Cancer Res.* **68**, 307–313 (2008).

132. Pett, M. & Coleman, N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J. Pathol.* **212**, 356–367 (2007).

133. Matsukura, T., Koi, S. & Sugase, M. Both episomal and integrated forms of human papillomavirus type 16 are involved in invasive cervical cancers. *Virology* **172**, 63–72 (1989).

134. Johannsen, E. & Lambert, P. F. Epigenetics of human papillomaviruses. *Virology* **445**, 205–212 (2013).

135. Yim, E.-K. & Park, J.-S. The Role of HPV E6 and E7 Oncoproteins in HPV-associated Cervical Carcinogenesis. *Cancer Res Treat* **37**, 319–324 (2005).

136. White, A. E., Livanos, E. M. & Tlsty, T. D. Differential disruption of genomic integrity and cell cycle regulation in normal human fibroblasts by the HPV oncoproteins. *Genes Dev.* **8**, 666–677 (1994).

137. Isaacson Wechsler, E. *et al.* Reconstruction of human papillomavirus type 16-mediated early-stage neoplasia implicates E6/E7 deregulation and the loss of contact inhibition in neoplastic progression. *J. Virol.* **86**, 6358–6364 (2012).

138. Münger, K. *et al.* Mechanisms of human papillomavirus-induced oncogenesis. *J. Virol.* **78**, 11451–11460 (2004).

139. Ojesina, A. I. *et al.* Landscape of genomic alterations in cervical carcinomas. *Nature* **506**, 371–375 (2014).

140. Cole, S. T. & Danos, O. Nucleotide sequence and comparative analysis of the human papillomavirus type 18 genome. Phylogeny of papillomaviruses and repeated structure of the E6 and E7 gene products. *J. Mol. Biol.* **193**, 599–608 (1987).

141. Barbosa, M. S. & Wettstein, F. O. Transcription of the cottontail rabbit papillomavirus early region and identification of two E6 polypeptides in COS-7 cells. *J Virol* **61**, 2938–2942 (1987).

142. Barbosa, M. S., Vass, W. C., Lowy, D. R. & Schiller, J. T. In vitro biological activities of the E6 and E7 genes vary among human papillomaviruses of different oncogenic potential. *J. Virol.* **65**, 292–298 (1991).

143. Patel, D., Huang, S. M., Baglia, L. A. & McCance, D. J. The E6 protein of human papillomavirus type 16 binds to and inhibits co-activation by CBP and p300. *EMBO J.* **18**, 5061–5072 (1999).

144. Jackson, S., Harwood, C., Thomas, M., Banks, L. & Storey, A. Role of Bak in UV-induced apoptosis in skin cancer and abrogation by HPV E6 proteins. *Genes Dev.* **14**, 3065–3073 (2000).

145. Thomas, M. & Banks, L. Inhibition of Bak-induced apoptosis by HPV-18 E6. *Oncogene* **17**, 2943–2954 (1998).

146. Underbrink, M. P., Howie, H. L., Bedard, K. M., Koop, J. I. & Galloway, D. A. E6 proteins from multiple human betapapillomavirus types degrade Bak and protect keratinocytes from apoptosis after UVB irradiation. *J. Virol.* **82**, 10408–10417 (2008).

147. Banks, L., Pim, D. & Thomas, M. Human tumour viruses and the deregulation of cell polarity in cancer. *Nat. Rev. Cancer* **12**, 877–886 (2012).

148. Borbély, A. A. *et al.* Effects of human papillomavirus type 16 oncoproteins on survivin gene expression. *J. Gen. Virol.* **87**, 287–294 (2006).

149. Filippova, M., Song, H., Connolly, J. L., Dermody, T. S. & Duerksen-Hughes, P. J. The human papillomavirus 16 E6 protein binds to tumor necrosis factor (TNF) R1 and protects cells from TNF-induced apoptosis. *J. Biol. Chem.* **277**, 21730–21739 (2002).

150. Filippova, M., Parkhurst, L. & Duerksen-Hughes, P. J. The human papillomavirus 16 E6 protein binds to Fas-associated death domain and protects cells from Fas-triggered apoptosis. *J. Biol. Chem.* **279**, 25729–25744 (2004).

151. Moody, C. A., Fradet-Turcotte, A., Archambault, J. & Laimins, L. A. Human papillomaviruses activate caspases upon epithelial differentiation to induce viral genome amplification. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19541–19546 (2007).

152. Manzo-Merino, J., Massimi, P., Lizano, M. & Banks, L. The human papillomavirus (HPV) E6 oncoproteins promotes nuclear localization of active caspase 8. *Virology* **450–451**, 146–152 (2014).

153. Rawls, J. A., Pusztai, R. & Green, M. Chemical synthesis of human papillomavirus type 16 E7 oncoprotein: autonomous protein domains for induction of cellular DNA synthesis and for trans activation. *J Virol* **64**, 6121–6129 (1990).

154. McIntyre, M. C., Frattini, M. G., Grossman, S. R. & Laimins, L. A. Human papillomavirus type 18 E7 protein requires intact Cys-X-X-Cys motifs for zinc binding, dimerization, and transformation but not for Rb binding. *J Virol* **67**, 3142–3150 (1993).

155. Banks, L., Edmonds, C. & Vousden, K. H. Ability of the HPV16 E7 protein to bind RB and induce DNA synthesis is not sufficient for efficient transforming activity in NIH3T3 cells. *Oncogene* **5**, 1383–1389 (1990).

156. Huh, K. *et al.* Human papillomavirus type 16 E7 oncoprotein associates with the cullin 2 ubiquitin ligase complex, which contributes to degradation of the retinoblastoma tumor suppressor. *J. Virol.* **81**, 9737–9747 (2007).

157. Funk, J. O. *et al.* Inhibition of CDK activity and PCNA-dependent DNA replication by p21 is blocked by interaction with the HPV-16 E7 oncoprotein. *Genes Dev.* **11**, 2090–2100 (1997).

158. Jones, D. L., Alani, R. M. & Münger, K. The human papillomavirus E7 oncoprotein can uncouple cellular differentiation and proliferation in human keratinocytes by abrogating p21Cip1-mediated inhibition of cdk2. *Genes Dev.* **11**, 2101–2111 (1997).

159. Zerfass-Thome, K. *et al.* Inactivation of the cdk inhibitor p27KIP1 by the human papillomavirus type 16 E7 oncoprotein. *Oncogene* **13**, 2323–2330 (1996).

160. zur Hausen, H. Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis. *J. Natl. Cancer Inst.* **92**, 690–698 (2000).

161. Chan, S. Y. *et al.* Phylogenetic analysis of 48 papillomavirus types and 28 subtypes and variants: a showcase for the molecular evolution of DNA viruses. *J. Virol.* **66**, 5714–5725 (1992).

162. Van Ranst, M., Kaplan, J. B. & Burk, R. D. Phylogenetic classification of human papillomaviruses: correlation with clinical manifestations. *J. Gen. Virol.* **73 (Pt 10)**, 2653–2660 (1992).

163. de Villiers, E. M. Human pathogenic papillomavirus types: an update. *Curr. Top. Microbiol. Immunol.* **186**, 1–12 (1994).

164. Chan, S. Y., Delius, H., Halpern, A. L. & Bernard, H. U. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. *J. Virol.* **69**, 3074–3083 (1995).

165. de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. & zur Hausen, H. Classification of papillomaviruses. *Virology* **324**, 17–27 (2004).

166. Van Regenmortel, M. H., Maniloff, J. & Calisher, C. The concept of virus species. *Arch. Virol.* **120**, 313–314 (1991).

167. C.M. Fauquet, J. Maniloff, L.A. Ball, U. Desselberger & M.A. Mayo. *Virus Taxonomy - Eighth Report of the International Committee on Taxonomy of Viruses*. (2005).

168. Bernard, H.-U. *et al.* Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401**, 70–79 (2010).

169. Bzhalava, D., Eklund, C. & Dillner, J. International standardization and classification of human papillomavirus types. *Virology* **476**, 341–344 (2015).

170. de Villiers, E.-M. Cross-roads in the classification of papillomaviruses. *Virology* **445**, 2–10 (2013).

171. Van Doorslaer, K. *et al.* The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.* **41**, D571-578 (2013).
172. Van Doorslaer, K. *et al.* The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res* **45**, D499–D506 (2017).
173. National center for biotechnology information U.S National Library of Medicine. *Entrez Sequences Help - NCBI help manual*. (2010).
174. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988]. National Center for Biotechnology Information (NCBI). (2020).
175. Egawa, N., Egawa, K., Griffin, H. & Doorbar, J. Human Papillomaviruses; Epithelial Tropisms, and the Development of Neoplasia. *Viruses* **7**, 3863–3890 (2015).
176. Halec, G. *et al.* Biological activity of probable/possible high-risk human papillomavirus types in cervical cancer. *Int. J. Cancer* **132**, 63–71 (2013).
177. Bouvard, V. *et al.* A review of human carcinogens--Part B: biological agents. *Lancet Oncol.* **10**, 321–322 (2009).
178. McLaughlin-Drubin, M. E., Meyers, J. & Munger, K. Cancer associated human papillomaviruses. *Curr Opin Virol* **2**, 459–466 (2012).
179. Bruggink, S. C. *et al.* Cutaneous wart-associated HPV types: prevalence and relation with patient characteristics. *J. Clin. Virol.* **55**, 250–255 (2012).
180. King, C. M. *et al.* Human papillomavirus types 2, 27, and 57 Identified in plantar verrucae from HIV-positive and HIV-negative individuals. *J Am Podiatr Med Assoc* **104**, 141–146 (2014).
181. Hogendoorn, G. K. *et al.* Morphological characteristics and human papillomavirus genotype predict the treatment response in cutaneous warts. *Br. J. Dermatol.* **178**, 253–260 (2018).
182. Antonsson, A., Forslund, O., Ekberg, H., Sterner, G. & Hansson, B. G. The Ubiquity and Impressive Genomic Diversity of Human Skin Papillomaviruses Suggest a Commensalic Nature of These Viruses. *J. Virol.* **74**, 11636–11641 (2000).

183. Tommasino, M. The biology of beta human papillomaviruses. *Virus Res.* **231**, 128–138 (2017).

184. Wong, M. C. S. *et al.* Prevalence and Epidemiologic Profile of Oral Infection with Alpha, Beta, and Gamma Papillomaviruses in an Asian Chinese Population. *J. Infect. Dis.* **218**, 388–397 (2018).

185. Nunes, E. M., Talpe-Nunes, V. & Sichero, L. Epidemiology and biology of cutaneous human papillomavirus. *Clinics (Sao Paulo)* **73**, (2018).

186. Hampras, S. S. *et al.* Prevalence and Concordance of Cutaneous Beta Human Papillomavirus Infection at Mucosal and Cutaneous Sites. *J. Infect. Dis.* **216**, 92–96 (2017).

187. Smelov, V. *et al.* Prevalence of cutaneous beta and gamma human papillomaviruses in the anal canal of men who have sex with women. *Papillomavirus Research* **3**, 66–72 (2017).

188. Forslund, O., Johansson, H., Madsen, K. G. & Kofoed, K. The Nasal Mucosa Contains a Large Spectrum of Human Papillomavirus Types from the Betapapillomavirus and Gammapapillomavirus Genera. *J Infect Dis* **208**, 1335–1341 (2013).

189. Cornet, I. *et al.* Comparative analysis of transforming properties of E6 and E7 from different beta human papillomavirus types. *J. Virol.* **86**, 2366–2370 (2012).

190. Viarisio, D. *et al.* Novel ß-HPV49 Transgenic Mouse Model of Upper Digestive Tract Cancer. *Cancer Res.* **76**, 4216–4225 (2016).

191. Bolatti, E. M. *et al.* Assessing Gammapapillomavirus infections of mucosal epithelia with two broad-spectrum PCR protocols. *BMC Infect. Dis.* **20**, 274 (2020).

192. Bottalico, D. *et al.* Characterization of human papillomavirus type 120: a novel betapapillomavirus with tropism for multiple anatomical niches. *J Gen Virol* **93**, 1774–1779 (2012).

193. Dunne, E. F. & Park, I. U. HPV and HPV-associated diseases. *Infect. Dis. Clin. North Am.* **27**, 765–778 (2013).

194. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975-2009, featuring the burden and trends in human papillomavirus(HPV)-associated cancers and HPV vaccination coverage levels. *J. Natl. Cancer Inst.* **105**, 175–201 (2013).

195. Palefsky, J. Human papillomavirus infection in HIV-infected persons. *Top HIV Med* **15**, 130–133 (2007).

196. Denny, L. A. *et al.* Human papillomavirus, human immunodeficiency virus and immunosuppression. *Vaccine* **30 Suppl 5**, F168-174 (2012).

197. Freiberger, D., Lewis, L. & Helfand, L. Human papillomavirus-related high-grade squamous intraepithelial lesions of the esophagus, skin, and cervix in an adolescent lung transplant recipient: a case report and literature review. *Transpl Infect Dis* **17**, 98–102 (2015).

198. Sias, C. *et al.* Alpha, Beta, gamma human PapillomaViruses (HPV) detection with a different sets of primers in oropharyngeal swabs, anal and cervical samples. *Virol J* **16**, (2019).

199. Forman, D. *et al.* Global burden of human papillomavirus and related diseases. *Vaccine* **30 Suppl 5**, F12-23 (2012).

200. Tommasino, M. The human papillomavirus family and its role in carcinogenesis. *Semin. Cancer Biol.* **26**, 13–21 (2014).

201. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Biological agents. Volume 100 B. A review of human carcinogens. *IARC Monogr Eval Carcinog Risks Hum* **100**, 1–441 (2012).

202. Rollison, D. E., Viarisio, D., Amorrortu, R. P., Gheit, T. & Tommasino, M. An Emerging Issue in Oncogenic Virology: the Role of Beta Human Papillomavirus Types in the Development of Cutaneous Squamous Cell Carcinoma. *J. Virol.* **93**, (2019).

203. Tommasino, M. HPV and skin carcinogenesis. *Papillomavirus Res* **7**, 129–131 (2019).

204. Lindelöf, B., Sigurgeirsson, B., Gäbel, H. & Stern, R. S. Incidence of skin cancer in 5356 patients following organ transplantation. *Br. J. Dermatol.* **143**, 513–519 (2000).

205. Neale, R. E. *et al.* Human papillomavirus load in eyebrow hair follicles and risk of cutaneous squamous cell carcinoma. *Cancer Epidemiol. Biomarkers Prev.* **22**, 719–727 (2013).

206. Bouwes Bavinck, J. N. *et al.* Human papillomavirus and posttransplantation cutaneous squamous cell carcinoma: A multicenter, prospective cohort study. *Am. J. Transplant.* **18**, 1220–1230 (2018).

207. Chahoud, J. *et al.* Association Between β-Genus Human Papillomavirus and Cutaneous Squamous Cell Carcinoma in Immunocompetent Individuals-A Meta-analysis. *JAMA Dermatol* **152**, 1354–1364 (2016).

208. Grace, M. & Munger, K. Proteomic analysis of the gamma human papillomavirus type 197 E6 and E7 associated cellular proteins. *Virology* **500**, 71–81 (2017).

209. Arroyo Mühr, L. S. *et al.* Human papillomavirus type 197 is commonly present in skin tumors. *Int. J. Cancer* **136**, 2546–2555 (2015).

210. Jadassohn, J. Verh. dtsch, derm. Ges. 5, 497. (1896).

211. Rowson, K. E. & Mahy, B. W. Human papova (wart) virus. *Bacteriol Rev* **31**, 110–131 (1967).

212. Guillet, G. Y., del Grande, P. & Thivolet, J. Cutaneous and mucosal warts. Clinical and histopathological criteria for classification. *Int. J. Dermatol.* **21**, 89–93 (1982).

213. Al Aboud, A. M. & Nigam, P. K. Wart (Plantar, Verruca Vulgaris, Verrucae). in *StatPearls* (StatPearls Publishing, 2020).

214. Hoy, T., Singhal, P. K., Willey, V. J. & Insinga, R. P. Assessing incidence and economic burden of genital warts with data from a US commercially insured population. *Curr Med Res Opin* **25**, 2343–2351 (2009).

215. Hartwig, S., St Guily, J. L., Dominiak-Felden, G., Alemany, L. & de Sanjosé, S. Estimation of the overall burden of cancers, precancerous lesions, and genital warts attributable to 9-valent HPV vaccine types in women and men in Europe. *Infect Agent Cancer* **12**, (2017).

216. Yanofsky, V. R., Patel, R. V. & Goldenberg, G. Genital Warts. *J Clin Aesthet Dermatol* **5**, 25–36 (2012).

217. Jabłońska S, Majewski S, Obalek S & Orth G. Cutaneous warts. *Clin Dermatol* 15(3):309-19. (1997).

218. Burd, E. M. & Dean, C. L. Human Papillomavirus. *Microbiol Spectr* **4**, (2016).

219. Wheless, L., Jacks, S., Mooneyham, K. A., Leach, B. C. & Cook, J. Skin cancer in organ transplant recipients: more than the immune system. *J Am Acad Dermatol* **71**, 359–365 (2014).

220. Bouwes Bavinck, J. N. & Berkhout, R. J. HPV infections and immunosuppression. *Clin. Dermatol.* **15**, 427–437 (1997).

221. Bouwes Bavinck, J. N., Feltkamp, M., Struijk, L. & ter Schegget, J. Human papillomavirus infection and skin cancer risk in organ transplant recipients. *J. Investig. Dermatol. Symp. Proc.* **6**, 207–211 (2001).

222. Greenspan D, de Villiers EM, Greenspan JS, de Souza YG & zur Hausen H. Unusual HPV types in oral warts in association with HIV infection. *J Oral Pathol* (9-10):482-8. (1988).

223. de Villiers EM. Prevalence of HPV 7 papillomas in the oral mucosa and facial skin of patients with human immunodeficiency virus. *Arch Dermatol* 125(11):1590. (1989).

224. Myers, D. J. & Fillman, E. P. Epidermodysplasia Verruciformis. in *StatPearls* (StatPearls Publishing, 2020).

225. Patel, T., Morrison, L. K., Rady, P. & Tyring, S. Epidermodysplasia verruciformis and susceptibility to HPV. *Dis. Markers* **29**, 199–206 (2010).

226. Rogers HD *et al.* Acquired epidermodysplasia verruciformis. *J Am Acad Dermatol* l 60:315–320 (2009).

227. Gül, U., Kiliç, A., Gönül, M., Cakmak, S. K. & Bayis, S. S. Clinical aspects of epidermodysplasia verruciformis and review of the literature. *Int. J. Dermatol.* **46**, 1069–1072 (2007).

228. de Oliveira, W. R. P., Festa Neto, C., Rady, P. L. & Tyring, S. K. Clinical aspects of epidermodysplasia verruciformis. *J Eur Acad Dermatol Venereol* **17**, 394–398 (2003).

229. Gewirtzman, A., Bartlett, B. & Tyring, S. Epidermodysplasia verruciformis and human papilloma virus. *Curr. Opin. Infect. Dis.* **21**, 141–146 (2008).

230. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359-386 (2015).

231. Wagner, S. *et al.* [HPV-associated oropharyngeal cancer-incidence, trends, diagnosis, and treatment]. *Urologe A* **57**, 1457–1463 (2018).

232. Marur, S., D'Souza, G., Westra, W. H. & Forastiere, A. A. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol.* **11**, 781–789 (2010).

233. Barul, C. *et al.* Occupational exposure to chlorinated solvents and risk of head and neck cancer in men: a population-based case-control study in France. *Environ Health* **16**, 77 (2017).

234. Stornetta, A., Guidolin, V. & Balbo, S. Correction: Alessia Stornetta et al. Alcohol-Derived Acetaldehyde Exposure in the Oral Cavity. Cancers 2018, 10, 20. *Cancers (Basel)* **10**, (2018).

235. Riaz, N. *et al.* A nomogram to predict loco-regional control after re-irradiation for head and neck cancer. *Radiother Oncol* **111**, 382–387 (2014).

236. Hille, J. J., Webster-Cyriaque, J., Palefski, J. M. & Raab-Traub, N. Mechanisms of expression of HHV8, EBV and HPV in selected HIV-associated oral lesions. *Oral Dis* **8 Suppl 2**, 161–168 (2002).

237. Mathur, S., Conway, D. I., Worlledge-Andrew, H., Macpherson, L. M. D. & Ross, A. J. Assessment and prevention of behavioural and social risk factors associated with oral cancer: protocol for a systematic review of clinical guidelines and systematic reviews to inform Primary Care dental professionals. *Syst Rev* **4**, (2015).

238. Fakhry, C. *et al.* Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. *J. Natl. Cancer Inst.* **100**, 261–269 (2008).

239. Gheit, T. *et al.* Role of mucosal high-risk human papillomavirus types in head and neck cancers in central India. *Int. J. Cancer* **141**, 143–151 (2017).

240. Sturgis, E. M. & Cinciripini, P. M. Trends in head and neck cancer incidence in relation to smoking prevalence: an emerging epidemic of human papillomavirus-associated cancers? *Cancer* **110**, 1429–1435 (2007).

241. Gillison, M. L. *et al.* Eurogin Roadmap: comparative epidemiology of HPV infection and associated cancers of the head and neck and cervix. *Int. J. Cancer* **134**, 497–507 (2014).

242. Viens, L. J. *et al.* Human Papillomavirus-Associated Cancers - United States, 2008-2012. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 661–666 (2016).

243. Pedro Diz *et al.* Oral and pharyngeal cancer in Europe: Incidence, mortality and trends as presented to the Global Oral Cancer Forum. *sage journals* (2017).

244. Nguyen, N. P. *et al.* Oral sex and oropharyngeal cancer. *Medicine (Baltimore)* **95**, (2016).

245. Klussmann, J. P. *et al.* Prevalence, distribution, and viral load of human papillomavirus 16 DNA in tonsillar carcinomas. *Cancer* **92**, 2875–2884 (2001).

246. Pai, S. I. & Westra, W. H. Molecular pathology of head and neck cancer: implications for diagnosis, prognosis, and treatment. *Annu Rev Pathol* **4**, 49–70 (2009).

247. Maruyama, H. *et al.* Human papillomavirus and p53 mutations in head and neck squamous cell carcinoma among Japanese population. *Cancer Sci* **105**, 409–417 (2014).

248. Buitrago-Pérez, Á., Garaulet, G., Vázquez-Carballo, A., Paramio, J. M. & García-Escudero, R. Molecular Signature of HPV-Induced Carcinogenesis: pRb, p53 and Gene Expression Profiling. *Curr Genomics* **10**, 26–34 (2009).

249. Nevens, D. & Nuyts, S. HPV-positive head and neck tumours, a distinct clinical entity. *B-ENT* **11**, 81–87 (2015).

250. Chu, A., Genden, E., Posner, M. & Sikora, A. A patient-centered approach to counseling patients with head and neck cancer undergoing human papillomavirus testing: a clinician's guide. *Oncologist* **18**, 180–189 (2013).

251. Cameron, J. E. & Hagensee, M. HPV-Associated Oropharyngeal Cancer in the HIV/AIDS Patient. *Cancer Treat. Res.* **177**, 131–181 (2019).

252. Stier, E. Human Papillomavirus Related Diseases in HIV-infected individuals. *Curr Opin Oncol* **20**, 541–546 (2008).

253. Randall T. Hayden, Donna M. Wolk, Karen C. Carroll & Yi-Wei Tang. *Diagnostic Microbiology of the Immunocompromised Host*. (2016).

254. LIU, G., SHARMA, M., TAN, N. & BARNABAS, R. HIV-positive women have higher risk of HPV infection, precancerous lesions, and cervical cancer: A systematic review and meta-analysis. *AIDS* **32**, 795–808 (2018).

255. Madeleine, M. M., Finch, J. L., Lynch, C. F., Goodman, M. T. & Engels, E. A. HPV-Related Cancers after Solid Organ Transplantation in the US. *Am J Transplant* **13**, 3202–3209 (2013).

256. Mellin, H. *et al.* Human papillomavirus type 16 is episomal and a high viral load may be correlated to better prognosis in tonsillar cancer. *Int. J. Cancer* **102**, 152–158 (2002).

257. Koskinen, W. J. *et al.* Prevalence and physical status of human papillomavirus in squamous cell carcinomas of the head and neck. *Int. J. Cancer* **107**, 401–406 (2003).

258. Johnson, L. G., Madeleine, M. M., Newcomer, L. M., Schwartz, S. M. & Daling, J. R. Anal cancer incidence and survival: the surveillance, epidemiology, and end results experience, 1973-2000. *Cancer* **101**, 281–288 (2004).

259. de Pokomandy, A. *et al.* HAART and progression to high-grade anal intraepithelial neoplasia in men who have sex with men and are infected with HIV. *Clin. Infect. Dis.* **52**, 1174–1181 (2011).

260. Silverberg, M. J. *et al.* Risk of anal cancer in HIV-infected and HIV-uninfected individuals in North America. *Clin. Infect. Dis.* **54**, 1026–1034 (2012).

261. Bertisch, B. *et al.* Risk factors for anal cancer in persons infected with HIV: a nested case-control study in the Swiss HIV Cohort Study. *Am. J. Epidemiol.* **178**, 877–884 (2013).

262. Machalek, D. A. *et al.* Anal human papillomavirus infection and associated neoplastic lesions in men who have sex with men: a systematic review and meta-analysis. *Lancet Oncol.* **13**, 487–500 (2012).

263. Robbins, H. A. *et al.* Excess cancers among HIV-infected people in the United States. *J. Natl. Cancer Inst.* **107**, (2015).

264. Tong, W. W. Y., Hillman, R. J., Kelleher, A. D., Grulich, A. E. & Carr, A. Anal intraepithelial neoplasia and squamous cell carcinoma in HIV-infected adults. *HIV Med.* **15**, 65–76 (2014).

265. Darwich, L. *et al.* Distribution of human papillomavirus genotypes in anal cytological and histological specimens from HIV-infected men who have sex with men and men who have sex with women. *Dis. Colon Rectum* **56**, 1043–1052 (2013).

266. Darragh, T. M. & Winkler, B. Anal cancer and cervical cancer screening: key differences. *Cancer Cytopathol* **119**, 5–19 (2011).

267. Darragh, T. M. *et al.* The Lower Anogenital Squamous Terminology Standardization Project for HPV-Associated Lesions: background and consensus recommendations from the College of American Pathologists and the American Society for Colposcopy and Cervical Pathology. *Arch. Pathol. Lab. Med.* **136**, 1266–1297 (2012).

268. Hillman, R. J. *et al.* 2016 IANS International Guidelines for Practice Standards in the Detection of Anal Cancer Precursors. *J Low Genit Tract Dis* **20**, 283–291 (2016).

269. Antinori, A. *et al.* Italian guidelines for the use of antiretroviral agents and the diagnostic-clinical management of HIV-1 infected persons. Update 2015. *New Microbiol.* **39**, 93–109 (2016).

270. Kalof, A. N. & Cooper, K. p16INK4a immunoexpression: surrogate marker of high-risk HPV and high-grade cervical intraepithelial neoplasia. *Adv Anat Pathol* **13**, 190–194 (2006).

271. Sarwath, H. *et al.* Introduction of p16INK4a as a surrogate biomarker for HPV in women with invasive cervical cancer in Sudan. *Infect Agent Cancer* **12**, (2017).

272. Sahasrabuddhe, V. V., Luhn, P. & Wentzensen, N. Human papillomavirus and cervical cancer: biomarkers for improved prevention efforts. *Future Microbiol* **6**, 1083–1098 (2011).

273. Klaes, R. *et al.* Overexpression of p16(INK4A) as a specific marker for dysplastic and neoplastic epithelial cells of the cervix uteri. *Int. J. Cancer* **92**, 276–284 (2001).

274. Donà, M. G. *et al.* Anal cytological lesions and HPV infection in individuals at increased risk for anal cancer. *Cancer Cytopathol* **126**, 461–470 (2018).

275. Hernandez, B. Y. *et al.* Burden of invasive squamous cell carcinoma of the penis in the United States, 1998-2003. *Cancer* **113**, 2883–2891 (2008).

276. Albersen, M. & Spiess, P. E. Editorial: penile cancer. *Curr Opin Urol* **29**, 143–144 (2019).

277. Hernandez, B. Y. *et al.* Human papillomavirus genotype prevalence in invasive penile cancers from a registry-based United States population. *Front Oncol* **4**, 9 (2014).

278. Backes, D. M., Kurman, R. J., Pimenta, J. M. & Smith, J. S. Systematic review of human papillomavirus prevalence in invasive penile cancer. *Cancer Causes Control* **20**, 449–457 (2009).

279. Gormley, R. H. & Kovarik, C. L. Dermatologic manifestations of HPV in HIV-infected individuals. *Curr HIV/AIDS Rep* **6**, 130–138 (2009).

280. Bleeker, M. C. G. *et al.* Penile cancer: epidemiology, pathogenesis and prevention. *World J Urol* **27**, 141–150 (2009).

281. Hakenberg, O. W. *et al.* The Diagnosis and Treatment of Penile Cancer. *Dtsch Arztebl Int* **115**, 646–652 (2018).

282. Mentrikoski, M. J., Stelow, E. B., Culp, S., Frierson, H. F. & Cathro, H. P. Histologic and immunohistochemical assessment of penile carcinomas in a North American population. *Am. J. Surg. Pathol.* **38**, 1340–1348 (2014).

283. Chaux, A. *et al.* Combining routine morphology, p16(INK4a) immunohistochemistry, and in situ hybridization for the detection of human papillomavirus infection in penile carcinomas: a

tissue microarray study using classifier performance analyses. *Urol. Oncol.* **32**, 171–177 (2014).

284. Tang, D. H. *et al.* Lack of P16ink4a over expression in penile squamous cell carcinoma is associated with recurrence after lymph node dissection. *J. Urol.* **193**, 519–525 (2015).
285. Schiffman, M. & Kjaer, S. K. Chapter 2: Natural history of anogenital human papillomavirus infection and neoplasia. *J. Natl. Cancer Inst. Monographs* 14–19 (2003).
286. Moscicki, A.-B., Schiffman, M., Kjaer, S. & Villa, L. L. Chapter 5: Updating the natural history of HPV and anogenital cancer. *Vaccine* **24 Suppl 3**, S3/42-51 (2006).
287. Moscicki, A.-B. *et al.* Updating the natural history of human papillomavirus and anogenital cancers. *Vaccine* **30 Suppl 5**, F24-33 (2012).
288. Schiffman, M. *et al.* The carcinogenicity of human papillomavirus types reflects viral evolution. *Virology* **337**, 76–84 (2005).
289. Byun, J. M. *et al.* Persistent HPV-16 infection leads to recurrence of high-grade cervical intraepithelial neoplasia. *Medicine (Baltimore)* **97**, (2018).
290. Bulk, S. *et al.* The contribution of HPV18 to cervical cancer is underestimated using high-grade CIN as a measure of screening efficiency. *Br J Cancer* **96**, 1234–1236 (2007).
291. Trottier, H. *et al.* Type-specific duration of human papillomavirus infection: implications for human papillomavirus screening and vaccination. *J. Infect. Dis.* **197**, 1436–1447 (2008).
292. Berrington de González, A., Green, J. & International Collaboration of Epidemiological Studies of Cervical Cancer. Comparison of risk factors for invasive squamous cell carcinoma and adenocarcinoma of the cervix: collaborative reanalysis of individual data on 8,097 women with squamous cell carcinoma and 1,374 women with adenocarcinoma from 12 epidemiological studies. *Int. J. Cancer* **120**, 885–891 (2007).
293. Dugué, P.-A., Rebolj, M., Garred, P. & Lynge, E. Immunosuppression and risk of cervical cancer. *Expert Rev Anticancer Ther* **13**, 29–42 (2013).

294. Madeleine, M. M., Finch, J. L., Lynch, C. F., Goodman, M. T. & Engels, E. A. HPV-related cancers after solid organ transplantation in the United States. *Am. J. Transplant.* **13**, 3202–3209 (2013).

295. Brickman, C. & Palefsky, J. M. Human papillomavirus in the HIV-infected host: epidemiology and pathogenesis in the antiretroviral era. *Curr HIV/AIDS Rep* **12**, 6–15 (2015).

296. Rizzo, J. D. *et al.* Solid cancers after allogeneic hematopoietic cell transplantation. *Blood* **113**, 1175–1183 (2009).

297. Wang, Y., Brinch, L., Jebsen, P., Tanbo, T. & Kirschner, R. A clinical study of cervical dysplasia in long-term survivors of allogeneic stem cell transplantation. *Biol. Blood Marrow Transplant.* **18**, 747–753 (2012).

298. Savani, B. N. *et al.* Increased Risk of Cervical Dysplasia in Long-Term Survivors of Allogeneic Stem Cell Transplantation—Implications for Screening and HPV Vaccination. *Biol Blood Marrow Transplant* **14**, 1072–1075 (2008).

299. Strickler, H. D. *et al.* Natural history and possible reactivation of human papillomavirus in human immunodeficiency virus-positive women. *J. Natl. Cancer Inst.* **97**, 577–586 (2005).

300. Denton, K. J. Liquid based cytology in cervical cancer screening. *BMJ* **335**, 1–2 (2007).

301. Gay, J. D., Donaldson, L. D. & Goellner, J. R. False-negative results in cervical cytologic studies. *Acta Cytol.* **29**, 1043–1046 (1985).

302. Pileggi, C., Flotta, D., Bianco, A., Nobile, C. G. A. & Pavia, M. Is HPV DNA testing specificity comparable to that of cytological testing in primary cervical cancer screening? Results of a meta-analysis of randomized controlled trials. *Int. J. Cancer* **135**, 166–177 (2014).

303. Whitlock, E. P. *et al.* Liquid-based cytology and human papillomavirus testing to screen for cervical cancer: a systematic review for the U.S. Preventive Services Task Force. *Ann. Intern. Med.* **155**, 687–697, W214-215 (2011).

304. Saslow, D. *et al.* American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *CA Cancer J Clin* **62**, 147–172 (2012).

305. Ronco, G. *et al.* Efficacy of human papillomavirus testing for the detection of invasive cervical cancers and cervical intraepithelial neoplasia: a randomised controlled trial. *Lancet Oncol.* **11**, 249–257 (2010).

306. Kitchener, H. C. *et al.* ARTISTIC: a randomised trial of human papillomavirus (HPV) testing in primary cervical screening. *Health Technol Assess* **13**, 1–150, iii–iv (2009).

307. Rijkaart, D. C. *et al.* Human papillomavirus testing for the detection of high-grade cervical intraepithelial neoplasia and cancer: final results of the POBASCAM randomised controlled trial. *Lancet Oncol.* **13**, 78–88 (2012).

308. Naucler, P. *et al.* Human papillomavirus and Papanicolaou tests to screen for cervical cancer. *N. Engl. J. Med.* **357**, 1589–1597 (2007).

309. Dockter, J. *et al.* Analytical characterization of the APTIMA HPV Assay. *J. Clin. Virol.* **45 Suppl 1**, S39-47 (2009).

310. Sauter, J. L. *et al.* Testing of integrated human papillomavirus mRNA decreases colposcopy referrals: could a change in human papillomavirus detection methodology lead to more cost-effective patient care? *Acta Cytol.* **58**, 162–166 (2014).

311. Rousseau MC, Trevisan A, Villa L, Rohan T & Franco E. Viral load as a predictor of HPV infection persistence among women in the Ludwig-McGill cohort study in Sao Paulo, Brazil. *Proceedings of the 19th International Papillomavirus Conference, Florianopolis* (2001).

312. Cuzick J, Franco E & Monsonego J. Viral load as a surrogate for persistence in cervical human papillomavirus infection. : *New developments in cervical cancer screening and prevention. Oxford. Blackwell,373– 8.* (1997).

313. Schlecht, N. F. *et al.* Viral load as a predictor of the risk of cervical intraepithelial neoplasia. *Int. J. Cancer* **103**, 519–524 (2003).

314. Arias-Pulido, H., Peyton, C. L., Joste, N. E., Vargas, H. & Wheeler, C. M. Human papillomavirus type 16 integration in cervical carcinoma in situ and in invasive cervical cancer. *J. Clin. Microbiol.* **44**, 1755–1762 (2006).

315. Fontaine, J. *et al.* High levels of HPV-16 DNA are associated with high-grade cervical lesions in women at risk or infected with HIV. *AIDS* **19**, 785–794 (2005).

316. Gravitt, P. E. *et al.* Reproducibility of HPV 16 and HPV 18 viral load quantitation using TaqMan real-time PCR assays. *J. Virol. Methods* **112**, 23–33 (2003).

317. Wentzensen, N., Schiffman, M., Palmer, T. & Arbyn, M. Triage of HPV positive women in cervical cancer screening. *J. Clin. Virol.* **76 Suppl 1**, S49–S55 (2016).

318. Han, J., Colditz, G. A. & Hunter, D. J. Risk factors for skin cancers: a nested case-control study within the Nurses' Health Study. *Int J Epidemiol* **35**, 1514–1521 (2006).

319. Fuente, M. J. *et al.* A prospective study of the incidence of skin cancer and its risk factors in a Spanish Mediterranean population of kidney transplant recipients. *Br. J. Dermatol.* **149**, 1221–1226 (2003).

320. Del Marmol, V. & Stratigos, A. J. Editorial: New Frontiers in Skin Cancer. *Curr Opin Oncol* **31**, 53 (2019).

321. Laikova, K. V. *et al.* Advances in the Understanding of Skin Cancer: Ultraviolet Radiation, Mutations, and Antisense Oligonucleotides as Anticancer Drugs. *Molecules* **24**, (2019).

322. Howley, P. M. & Pfister, H. J. Beta genus papillomaviruses and skin cancer. *Virology* **479–480**, 290–296 (2015).

323. Lutzner, M. A. Epidermodysplasia verruciformis. An autosomal recessive disease characterized by viral warts and skin cancer. A model for viral oncogenesis. *Bull Cancer* **65**, 169–182 (1978).

324. Euvrard, S., Kanitakis, J. & Claudy, A. Skin cancers after organ transplantation. *N. Engl. J. Med.* **348**, 1681–1691 (2003).

325. O'Reilly Zwald, F. & Brown, M. Skin cancer in solid organ transplant recipients: advances in therapy and management: part I. Epidemiology of skin cancer in solid organ transplant recipients. *J. Am. Acad. Dermatol.* **65**, 253–261 (2011).

326. Nindl, I. & Rösl, F. Molecular concepts of virus infections causing skin cancer in organ transplant recipients. *Am. J. Transplant.* **8**, 2199–2204 (2008).

327. Terhorst, D., Drecoll, U., Stockfleth, E. & Ulrich, C. Organ transplant recipients and skin cancer: assessment of risk factors with focus on sun exposure. *Br. J. Dermatol.* **161 Suppl 3**, 85–89 (2009).

328. Weissenborn, S. *et al.* Beta-papillomavirus DNA loads in hair follicles of immunocompetent people and organ transplant recipients. *Med. Microbiol. Immunol.* **201**, 117–125 (2012).

329. Proby, C. M. *et al.* A case-control study of betapapillomavirus infection and cutaneous squamous cell carcinoma in organ transplant recipients. *Am. J. Transplant.* **11**, 1498–1508 (2011).

330. Andersson, K. *et al.* Prospective study of human papillomavirus seropositivity and risk of nonmelanoma skin cancer. *Am. J. Epidemiol.* **175**, 685–695 (2012).

331. Silverberg, M. J. *et al.* HIV infection status, immunodeficiency, and the incidence of non-melanoma skin cancer. *J. Natl. Cancer Inst.* **105**, 350–360 (2013).

332. Asgari, M. M., Ray, G. T., Quesenberry, C. P., Katz, K. A. & Silverberg, M. J. Association of Multiple Primary Skin Cancers With Human Immunodeficiency Virus Infection, CD4 Count, and Viral Load. *JAMA Dermatol* **153**, 892–896 (2017).

333. Grulich, A. E., van Leeuwen, M. T., Falster, M. O. & Vajdic, C. M. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet* **370**, 59–67 (2007).

334. Weissenborn, S. J. *et al.* Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers. *J. Invest. Dermatol.* **125**, 93–97 (2005).

335. Farzan, S. F. *et al.* Cutaneous alpha, beta and gamma human papillomaviruses in relation to squamous cell carcinoma of the skin: a population-based study. *Int. J. Cancer* **133**, 1713–1720 (2013).

336. Iannacone, M. R. *et al.* Case-control study of genus-beta human papillomaviruses in plucked eyebrow hairs and cutaneous squamous cell carcinoma. *Int. J. Cancer* **134**, 2231–2244 (2014).

337. Waterboer, T. *et al.* Serological association of beta and gamma human papillomaviruses with squamous cell carcinoma of the skin. *Br. J. Dermatol.* **159**, 457–459 (2008).

338. Forslund, O. *et al.* Cutaneous human papillomaviruses found in sun-exposed skin: Beta-papillomavirus species 2 predominates in squamous cell carcinoma. *J. Infect. Dis.* **196**, 876–883 (2007).

339. Bouwes Bavinck, J. N. *et al.* Multicenter study of the association between betapapillomavirus infection and cutaneous squamous cell carcinoma. *Cancer Res.* **70**, 9777–9786 (2010).

340. Karagas, M. R. *et al.* Genus beta human papillomaviruses and incidence of basal cell and squamous cell carcinomas of skin: population based case-control study. *BMJ* **341**, c2986 (2010).

341. Iannacone, M. R. *et al.* Case-control study of cutaneous human papillomaviruses in squamous cell carcinoma of the skin. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1303–1313 (2012).

342. Ally, M. S., Tang, J. Y. & Arron, S. T. Cutaneous human papillomavirus infection and basal cell carcinoma of the skin. *J Invest Dermatol* **133**, (2013).

343. Viarisio, D. *et al.* Beta HPV38 oncoproteins act with a hit-and-run mechanism in ultraviolet radiation-induced skin carcinogenesis in mice. *PLoS Pathog.* **14**, e1006783 (2018).

344. Caldeira, S. *et al.* The E6 and E7 proteins of the cutaneous human papillomavirus type 38 display transforming properties. *J. Virol.* **77**, 2195–2206 (2003).

345. Accardi, R. *et al.* Skin human papillomavirus type 38 alters p53 functions by accumulation of deltaNp73. *EMBO Rep.* **7**, 334–340 (2006).

346. Brimer, N., Lyons, C., Wallberg, A. E. & Vande Pol, S. B. Cutaneous papillomavirus E6 oncoproteins associate with MAML1 to repress transactivation and NOTCH signaling. *Oncogene* **31**, 4639–4646 (2012).

347. Tan, M. J. A. *et al.* Cutaneous β-human papillomavirus E6 proteins bind Mastermind-like coactivators and repress Notch signaling. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1473-1480 (2012).

348. Wu, L. *et al.* MAML1, a human homologue of Drosophila mastermind, is a transcriptional co-activator for NOTCH receptors. *Nat. Genet.* **26**, 484–489 (2000).

349. Meyers, J. M., Spangle, J. M. & Munger, K. The human papillomavirus type 8 E6 protein interferes with NOTCH activation during keratinocyte differentiation. *J. Virol.* **87**, 4762–4767 (2013).

350. Marcuzzi, G. P. *et al.* Spontaneous tumour development in human papillomavirus type 8 E6 transgenic mice and rapid induction by UV-light exposure and wounding. *J. Gen. Virol.* **90**, 2855–2864 (2009).

351. Hasche, D. *et al.* The interplay of UV and cutaneous papillomavirus infection in skin cancer development. *PLoS Pathog.* **13**, e1006723 (2017).

352. Van Doorslaer, K. *et al.* The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Res* **45**, D499–D506 (2017).

353. Guerrero, E. *et al.* Comparison of ViraPap, Southern hybridization, and polymerase chain reaction methods for human papillomavirus identification in an epidemiological investigation of cervical cancer. *J. Clin. Microbiol.* **30**, 2951–2959 (1992).

354. Kuypers, J. M. *et al.* Comparison of dot filter hybridization, Southern transfer hybridization, and polymerase chain reaction amplification for diagnosis of anal human papillomavirus infection. *J. Clin. Microbiol.* **31**, 1003–1006 (1993).

355. Morris, B. J. *et al.* Automated polymerase chain reaction for papillomavirus screening of cervicovaginal lavages: comparison with dot-blot hybridization in a sexually transmitted diseases clinic population. *J. Med. Virol.* **32**, 22–30 (1990).

356. Ward, P., Parry, G. N., Yule, R., Coleman, D. V. & Malcolm, A. D. Comparison between the polymerase chain reaction and slot blot hybridization for the detection of HPV sequences in cervical scrapes. *Cytopathology* **1**, 19–23 (1990).

357. Van den Brule AJ, Snijders P, Walboomers JM & Meijer CJ. PCR-based detection of genital HPV genotypes: an update and future perspectives. *Papillomavirus Report* **(4):95-9.**, (1993).

358. Hildesheim, A. *et al.* Persistence of type-specific human papillomavirus infection among cytologically normal women. *J. Infect. Dis.* **169**, 235–240 (1994).

359. MM Manos *et al.* The use of polymerase chain reaction amplification for the detection of genital human papillomaviruses , , , , , , M. Manos,. *Cancer Cell* **7:209–214**, (1989).

360. SNLIDERS, P. J. F. *et al.* The use of general primers in the polymerase chain reaction permits the detection of a broad spectrum of human papillomavirus genotypes. *Journal of General Virology* 71, 173-181. (1990).

361. Qu, W. *et al.* PCR detection of human papillomavirus: comparison between MY09/MY11 and GP5+/GP6+ primer systems. *J Clin Microbiol* **35**, 1304–1310 (1997).

362. van den Brule, A. J. *et al.* General primer-mediated polymerase chain reaction permits the detection of sequenced and still unsequenced human papillomavirus genotypes in cervical scrapes and carcinomas. *Int. J. Cancer* **45**, 644–649 (1990).

363. de Roda Husman, A. M. *et al.* Analysis of cytomorphologically abnormal cervical scrapes for the presence of 27 mucosotropic human papillomavirus genotypes, using polymerase chain reaction. *Int. J. Cancer* **56**, 802–806 (1994).

364. Shamanin, V., Delius, H. & de Villiers, E.-M. Development of a Broad Spectrum PCR Assay for Papillomaviruses and its Application in Screening Lung Cancer Biopsies. *Journal of General Virology,* **75**, 1149–1156 (1994).

365. Snijders, P. J., Meijer, C. J. & Walboomers, J. M. Degenerate primers based on highly conserved regions of amino acid sequence in papillomaviruses can be used in a generalized polymerase chain reaction to detect productive human papillomavirus infection. *J. Gen. Virol.* **72 (Pt 11)**, 2781–2786 (1991).

366. Ylitalo, N., Bergström, T. & Gyllensten, U. Detection of genital human papillomavirus by single-tube nested PCR and type-specific oligonucleotide hybridization. *J Clin Microbiol* **33**, 1822–1828 (1995).

367. Berkhout, R. J. *et al.* Nested PCR approach for detection and typing of epidermodysplasia verruciformis-associated human papillomavirus types in cutaneous cancers from renal transplant recipients. *J. Clin. Microbiol.* **33**, 690–695 (1995).

368. Boxman, I. L. *et al.* Detection of human papillomavirus DNA in plucked hairs from renal transplant recipients and healthy volunteers. *J. Invest. Dermatol.* **108**, 712–715 (1997).

369. van den Brule, A. J. *et al.* General primer polymerase chain reaction in combination with sequence analysis for identification of potentially novel human papillomavirus genotypes in cervical lesions. *J. Clin. Microbiol.* **30**, 1716–1721 (1992).

370. de Roda Husman, A. M., Walboomers, J. M., van den Brule, A. J., Meijer, C. J. & Snijders, P. J. The use of general primers GP5 and GP6 elongated at their 3' ends with adjacent highly conserved sequences improves human papillomavirus detection by PCR. *J. Gen. Virol.* **76 (Pt 4)**, 1057–1062 (1995).

371. Shamanin, V. *et al.* Human papillomavirus infections in nonmelanoma skin cancers from renal transplant recipients and nonimmunosuppressed patients. *J. Natl. Cancer Inst.* **88**, 802–811 (1996).

372. Barr, B. B. *et al.* Human papilloma virus infection and skin cancer in renal allograft recipients. *Lancet* **1**, 124–129 (1989).

373. Stark, L. A. *et al.* Prevalence of human papillomavirus DNA in cutaneous neoplasms from renal allograft recipients supports a possible viral role in tumour promotion. *Br. J. Cancer* **69**, 222–229 (1994).

374. de Villiers, E. M., Lavergne, D., McLaren, K. & Benton, E. C. Prevailing papillomavirus types in non-melanoma carcinomas of the skin in renal allograft recipients. *Int. J. Cancer* **73**, 356–361 (1997).

375. Gravitt, P. E. *et al.* Improved Amplification of Genital Human Papillomaviruses. *J Clin Microbiol* **38**, 357–361 (2000).

376. Baines, J. E., McGovern, R. M., Persing, D. & Gostout, B. S. Consensus-degenerate hybrid oligonucleotide primers (CODEHOP) for the detection of novel papillomaviruses and their application to esophageal and tonsillar carcinomas. *J. Virol. Methods* **123**, 81–87 (2005).

377. Los Alamos National Laboratory Bioscience Division. (2002).

378. Chouhy, D. *et al.* Identification of human papillomavirus type 156, the prototype of a new human gammapapillomavirus species, by a generic and highly sensitive PCR strategy for long DNA fragments. *Journal of General Virology,* **94**, 524–533 (2013).

379. Schmitt, M., Dondog, B., Waterboer, T. & Pawlita, M. Homogeneous amplification of genital human alpha papillomaviruses by PCR using novel broad-spectrum GP5+ and GP6+ primers. *J. Clin. Microbiol.* **46**, 1050–1059 (2008).

380. García, D. A. *et al.* Highly Sensitive Detection and Genotyping of HPV by PCR Multiplex and Luminex Technology in a Cohort of Colombian Women with Abnormal Cytology. *Open Virol J* **5**, 70–79 (2011).

381. Schmitt, M. *et al.* Bead-based multiplex genotyping of human papillomaviruses. *J. Clin. Microbiol.* **44**, 504–512 (2006).

382. Forslund, O., Antonsson, A., Nordin, P., Stenquist, B. & Hansson, B. G. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *J. Gen. Virol.* **80 (Pt 9)**, 2437–2443 (1999).

383. Chouhy, D. *et al.* New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. *Virology* **397**, 205–216 (2010).

384. Antonsson, A. & Hansson, B. G. Healthy Skin of Many Animal Species Harbors Papillomaviruses Which Are Closely Related to Their Human Counterparts. *J Virol* **76**, 12537–12542 (2002).

385. Forslund, O., Ly, H. & Higgins, G. Improved detection of cutaneous human papillomavirus DNA by single tube nested 'hanging droplet' PCR. *J. Virol. Methods* **110**, 129–136 (2003).

386. Walsh, E. E., Falsey, A. R., Swinburne, I. A. & Formica, M. A. Reverse transcription polymerase chain reaction (RT-PCR) for diagnosis of respiratory syncytial virus infection in adults: use of a single-tube 'hanging droplet' nested PCR. *J. Med. Virol.* **63**, 259–263 (2001).

387. Myers, G., G. M., C. Baker, K. Münger, F. Sverdrup, A. McBride &. H. U. Bernard. Alignments. In Human Papillomaviruses 1996. HPV Sequence Database. II-L1–1–67 (1996).

388. Munday, J. S., Tucker, R. S., Kiupel, M. & Harvey, C. J. Multiple oral carcinomas associated with a novel papillomavirus in a dog. *J. Vet. Diagn. Invest.* **27**, 221–225 (2015).

389. Kocjan, B. J., Hošnjak, L., Račnik, J., Zadravec, M. & Poljak, M. Complete Genome Sequence of Phodopus sungorus Papillomavirus Type 1 (PsPV1), a Novel Member of the Pipapillomavirus Genus, Isolated from a Siberian Hamster. *Genome Announc* **2**, (2014).

390. Stevens, H. *et al.* Complete Genome Sequence of the Crocuta crocuta Papillomavirus Type 1 (CcrPV1) from a Spotted Hyena, the First Papillomavirus Characterized in a Member of the Hyaenidae. *Genome Announc* **1**, (2013).

391. Ure, A. E. *et al.* Characterization of the complete genomes of Camelus dromedarius papillomavirus types 1 and 2. *J. Gen. Virol.* **92**, 1769–1777 (2011).

392. Claus, M. P. *et al.* Identification of unreported putative new bovine papillomavirus types in Brazilian cattle herds. *Vet. Microbiol.* **132**, 396–401 (2008).

393. Alotaibi, L., Provost, N., Gagnon, S., Franco, E. L. & Coutlée, F. Diversity of cutaneous human papillomavirus types in individuals with and without skin lesion. *J. Clin. Virol.* **36**, 133–140 (2006).

394. Nordin, P. *et al.* Human papilloma virus in skin, mouth and uterine cervix in female renal transplant recipients with or without a history of cutaneous squamous cell carcinoma. *Acta Derm. Venereol.* **87**, 219–222 (2007).

395. Forslund, O. *et al.* High prevalence of cutaneous human papillomavirus DNA on the top of skin tumors but not in 'Stripped' biopsies from the same tumors. *J. Invest. Dermatol.* **123**, 388–394 (2004).

396. Kocjan, B. J., Bzhalava, D., Forslund, O., Dillner, J. & Poljak, M. Molecular methods for identification and characterization of novel papillomaviruses. *Clin. Microbiol. Infect.* **21**, 808–816 (2015).

397. Forslund, O., DeAngelis, P. M., Beigi, M., Schjølberg, A. R. & Clausen, O. P. F. Identification of human papillomavirus in keratoacanthomas. *J. Cutan. Pathol.* **30**, 423–429 (2003).

398. Li, J. *et al.* Improved detection of human papillomavirus harbored in healthy skin with FAP6085/64 primers. *J. Virol. Methods* **193**, 633–638 (2013).

399. Gheit, T. *et al.* Development of a sensitive and specific multiplex PCR method combined with DNA microarray primer extension to detect Betapapillomavirus types. *J. Clin. Microbiol.* **45**, 2537–2544 (2007).

400. Ekström, J., Bzhalava, D., Svenback, D., Forslund, O. & Dillner, J. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. *Int. J. Cancer* **129**, 2643–2650 (2011).

401. Ekström, J. *et al.* Diversity of human papillomaviruses in skin lesions. *Virology* **447**, 300–311 (2013).

402. Bolatti, E. M. *et al.* High prevalence of Gammapapillomaviruses (Gamma-PVs) in pre-malignant cutaneous lesions of immunocompetent individuals using a new broad-spectrum

primer system, and identification of HPV210, a novel Gamma-PV type. *Virology* **525**, 182–191 (2018).

403. Brancaccio, R. N. *et al.* Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. *Virology* **520**, 1–10 (2018).
404. Rector, A., Tachezy, R. & Van Ranst, M. A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. *J. Virol.* **78**, 4993–4998 (2004).
405. Zobel, T., Iftner, T. & Stubenrauch, F. The Papillomavirus E8∧E2C Protein Represses DNA Replication from Extrachromosomal Origins. *Mol Cell Biol* **23**, 8352–8362 (2003).
406. Esteban, J. A., Salas, M. & Blanco, L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J. Biol. Chem.* **268**, 2719–2726 (1993).
407. Marincevic-Zuniga, Y., Gustavsson, I. & Gyllensten, U. Multiply-primed rolling circle amplification of human papillomavirus using sequence-specific primers. *Virology* **432**, 57–62 (2012).
408. Johne, R., Müller, H., Rector, A., van Ranst, M. & Stevens, H. Rolling-circle amplification of viral DNA genomes using phi29 polymerase. *Trends Microbiol.* **17**, 205–211 (2009).
409. Bzhalava, D. *et al.* Deep sequencing extends the diversity of human papillomaviruses in human skin. *Sci Rep* **4**, 5807 (2014).
410. Pray, L. Discovery of DNA structure and function: Watson and Crick. *Nature Education* 1(1):100 (2008).
411. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
412. Zallen, D. T. Despite Franklin's work, Wilkins earned his Nobel. *Nature* **425**, 15 (2003).
413. Hutchison, C. A. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* **35**, 6227–6237 (2007).

414. Holley, R. W. *et al.* STRUCTURE OF A RIBONUCLEIC ACID. *Science* **147**, 1462–1465 (1965).

415. Sanger, F., Brownlee, G. G. & Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* **13**, 373–398 (1965).

416. Brownlee, G. G. & Sanger, F. Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli. *J. Mol. Biol.* **23**, 337–353 (1967).

417. Cory, S., Marcker, K. A., Dube, S. K. & Clark, B. F. Primary structure of a methionine transfer RNA from Escherichia coli. *Nature* **220**, 1039–1040 (1968).

418. Dube, S. K. & Marcker, K. A. The nucleotide sequence of N-formyl-methionyl-transfer RNA. Partial digestion with pancreatic and T-1 ribonuclease and derivation of the total primary structure. *Eur. J. Biochem.* **8**, 256–262 (1969).

419. Goodman, H. M., Abelson, J., Landy, A., Brenner, S. & Smith, J. D. Amber suppression: a nucleotide change in the anticodon of a tyrosine transfer RNA. *Nature* **217**, 1019–1024 (1968).

420. Adams, J. M., Jeppesen, P. G., Sanger, F. & Barrell, B. G. Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA. *Nature* **223**, 1009–1014 (1969).

421. Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**, 82–88 (1972).

422. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).

423. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).

424. Chidgeavadze, Z. G. *et al.* 2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Res* **12**, 1671–1686 (1984).

425. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

426. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13**, 2399–2412 (1985).

427. Ansorge, W., Sproat, B. S., Stegemann, J. & Schwager, C. A non-radioactive automated method for DNA sequence determination. *J. Biochem. Biophys. Methods* **13**, 315–323 (1986).

428. Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. & Zenke, M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.* **15**, 4593–4602 (1987).

429. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).

430. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **18**, 1415–1419 (1990).

431. Luckey, J. A. *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* **18**, 4417–4421 (1990).

432. Hunkapiller, T., Kaiser, R. J., Koop, B. F. & Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 59–67 (1991).

433. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300**, 286–290 (2003).

434. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**, 2601–2610 (1979).

435. Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* **9**, 3015–3027 (1981).

436. Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).

437. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).

438. Jackson, D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* **69**, 2904–2909 (1972).

439. Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3240–3244 (1973).

440. Ansorge, W. J. Next-generation DNA sequencing techniques. *N Biotechnol* **25**, 195–203 (2009).

441. Nyrén, P. & Lundin, A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal. Biochem.* **151**, 504–509 (1985).

442. Hyman, E. D. A new method of sequencing DNA. *Anal. Biochem.* **174**, 423–436 (1988).

443. Simner, P. J., Khare, R. & Wengenack, N. L. Chapter 95 - Rapidly Growing Mycobacteria. in *Molecular Medical Microbiology (Second Edition)* (eds. Tang, Y.-W., Sussman, M., Liu, D., Poxton, I. & Schwartzman, J.) 1679–1690 (Academic Press, 2015).

444. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363, 365 (1998).

445. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

446. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).

447. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).

448. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* **55**, 641–658 (2009).

449. Davies K. 13 years ago, a beer summit in an English pub led to the birth of Solexa. **BioIT World**, (2010).

450. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

451. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**, 759–769 (2011).

452. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).

453. HiSeq X Ten Series of Sequencing Systems. *Illumina Official website* http://www.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf (2014).

454. An introduction to Next-Generation Sequencing Technology. *Illumina Official website* https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf (2017).

455. TUFTS - TUCF Genomics. http://tucf-genomics.tufts.edu/home/ordering.

456. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).

457. Bentley, D. R. *et al.* Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* **456**, 53–59 (2008).

458. Nakazato, T., Ohta, T. & Bono, H. Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive. *PLoS One* **8**, (2013).

459. Balasubramanian, S. Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun. (Camb.)* **47**, 7281–7286 (2011).

460. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

461. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).

462. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* **36**, e25 (2008).

463. Why does the per base sequence quality decrease over the read in Illumina? https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina.

464. Sanhueza, D., Guégan, J.-F., Jordan, H. & Chevillon, C. Environmental Variations in Mycobacterium ulcerans Transcriptome: Absence of Mycolactone Expression in Suboptimal Environments. *Toxins (Basel)* **11**, 146 (2019).

465. McEllistrem, M. C. Genetic diversity of the pneumococcal capsule: implications for molecular-based serotyping. *Future Microbiol* **4**, 857–865 (2009).

466. Lo, Y. M. D. & Chiu, R. W. K. Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. *Clin. Chem.* **55**, 607–608 (2009).

467. Ram, J. L., Karim, A. S., Sendler, E. D. & Kato, I. Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the Illumina sequencing platform. *Syst Biol Reprod Med* **57**, 162–170 (2011).

468. Yi, X. *et al.* Development and validation of a new HPV genotyping assay based on next-generation sequencing. *Am. J. Clin. Pathol.* **141**, 796–804 (2014).

469. Arroyo, L. S. *et al.* Next generation sequencing for human papillomavirus genotyping. *J. Clin. Virol.* **58**, 437–442 (2013).

470. Ambulos, N. P. *et al.* Next-Generation Sequencing-Based HPV Genotyping Assay Validated in Formalin-Fixed, Paraffin-Embedded Oropharyngeal and Cervical Cancer Specimens. *J Biomol Tech* **27**, 46–52 (2016).

471. Arroyo, L. S. *et al.* Next generation sequencing for human papillomavirus genotyping. *J. Clin. Virol.* **58**, 437–442 (2013).

472. Barzon, L. *et al.* Human papillomavirus genotyping by 454 next generation sequencing technology. *J. Clin. Virol.* **52**, 93–97 (2011).

473. Juliet Dang. Identification and characterization of novel human papillomaviruses in oral rinse samples. (2015).

474. Tuna, M. & Amos, C. I. Next generation sequencing and its applications in HPV-associated cancers. *Oncotarget* **8**, 8877–8889 (2017).

475. Speel, E. J. M. HPV Integration in Head and Neck Squamous Cell Carcinomas: Cause and Consequence. *Recent Results Cancer Res.* **206**, 57–72 (2017).

476. Meisal, R. *et al.* HPV Genotyping of Modified General Primer-Amplicons Is More Analytically Sensitive and Specific by Sequencing than by Hybridization. *PLoS ONE* **12**, e0169074 (2017).

477. Pastrana, D. V. *et al.* Metagenomic Discovery of 83 New Human Papillomavirus Types in Patients with Immunodeficiency. *mSphere* **3**, (2018).

478. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

479. Zhang, Q. *et al.* Combined immunodeficiency associated with DOCK8 mutations. *N. Engl. J. Med.* **361**, 2046–2055 (2009).

480. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

481. Tirosh, O. *et al.* Expanded skin virome in DOCK8-deficient patients. *Nat. Med.* **24**, 1815–1821 (2018).

482. Chevreux B, Wetter T & Suhai S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information;The German Conference on Bioinformatics. *Computer Science and Biology* (1999).

483. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

484. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

485. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3960–3964 (2003).

486. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).

487. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227-240 (2010).

488. Karamitros, T. *et al.* De Novo Assembly of Human Herpes Virus Type 1 (HHV-1) Genome, Mining of Non-Canonical Structures and Detection of Novel Drug-Resistance Mutations Using Short- and Long-Read Next Generation Sequencing Technologies. *PLoS ONE* **11**, e0157600 (2016).

489. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).

490. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).

491. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13770–13773 (1996).

492. Mardis, E. R. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)* **6**, 287–303 (2013).

493. Brown, C. G. & Clarke, J. Nanopore development at Oxford Nanopore. *Nat. Biotechnol.* **34**, 810–811 (2016).

494. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).

495. Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics* **13**, 4–16 (2015).

496. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).

497. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat Methods* **13**, 751–754 (2016).

498. Eisenstein, M. An ace in the hole for DNA sequencing. *Nature* **550**, 285–288 (2017).

499. Oxford Nanopore. https://nanoporetech.com/.

500. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).

501. Vanmechelen, B. *et al.* Identification of a novel species of papillomavirus in giraffe lesions using nanopore sequencing. *Vet. Microbiol.* **201**, 26–31 (2017).

502. Koren, S, Walenz, B.P., Berlin, K., Miller, J.R. & Phillippy, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv* (2016).

503. Quan, L. *et al.* Simultaneous detection and comprehensive analysis of HPV and microbiome status of a cervical liquid-based cytology sample using Nanopore MinION sequencing. *Sci Rep* **9**, (2019).

504. Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief. Bioinformatics* **20**, 1981–1996 (2019).

505. Moody G. Digital Code of Life: How Bioinformatics is revolutionizing Science, Medicine, and Business. *London: Wiley* (2004).

506. Pauling L & Zuckerkandl E. Chemical paleogenetics: molecular “restoration studies” of extinct forms of life. *Acta Chem Scand* 17:S9–16. (1963).

507. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

508. Murata, M., Richardson, J. S. & Sussman, J. L. Simultaneous comparison of three protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 3073–3077 (1985).

509. Feng, D. F. & Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360 (1987).

510. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).

511. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079**, 105–116 (2014).

512. Hert, D. G., Fredlake, C. P. & Barron, A. E. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**, 4618–4626 (2008).

513. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).

514. Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311 (1996).

515. Nascimento, F. F., Reis, M. D. & Yang, Z. A biologist’s guide to Bayesian phylogenetic analysis. *Nat Ecol Evol* **1**, 1446–1454 (2017).

516. Sheppard D. Beginner’s Introduction to Perl. *Perl.com* (2000).

517. Sharma V. Programming languages. In: Text Book of Bioinformatics. *Rastogi Publications* Chapter 5 (2008).

518. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).

519. Ekmekci, B., McAnany, C. E. & Mura, C. An Introduction to Programming for Bioscientists: A Python-Based Primer. *PLoS Comput. Biol.* **12**, e1004867 (2016).

520. Stoehr, P. J. & Cameron, G. N. The EMBL data library. *Nucleic Acids Res* **19**, 2227–2230 (1991).

521. Benson D, Lipman DJ & Ostell J. GenBank. *Nucleic Acids Res* 21:2963–5. (1993).

522. Li, Y. & Chen, L. Big biological data: challenges and opportunities. *Genomics Proteomics Bioinformatics* **12**, 187–189 (2014).

523. Gramates, L. S. *et al.* FlyBase at 25: looking to the future. *Nucleic Acids Res.* **45**, D663–D671 (2017).

524. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700-705 (2012).

525. Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* **46**, D762–D769 (2018).

526. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).

527. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39**, D28-31 (2011).

528. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).

529. Li, J.-W. *et al.* SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics* **28**, 1272–1273 (2012).

530. Parnell, L. D. *et al.* BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comput. Biol.* **7**, e1002216 (2011).

531. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767–1771 (2010).

532. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).

533. Andrews,S. FastQC: a quality control tool for high throughput sequence data. (2010).

534. PoreCamp. PoreCamp2016 Course Material : Understanding your MinION data. https://porecamp.github.io/2016/tutorials/PoreCamp2016-02-MinIONData.pdf (2016).

535. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

536. Adrien Leger and Tommaso Leonardi. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *JOSS* (2019).

537. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

538. Krueger F. Trim Galore!: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. (2015).

539. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J* **6**, 898–901 (2012).

540. Mende, D. R. *et al.* Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* **7**, e31386 (2012).

541. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315–327 (2010).

542. van der Walt, A. J. *et al.* Assembling metagenomes, one community at a time. *BMC Genomics* **18**, 521 (2017).

543. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

544. Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **12**, e1004957 (2016).

545. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

546. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).

547. Chen, Y. *et al.* VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**, 266–267 (2013).

548. Wang, Q., Jia, P. & Zhao, Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med* **7**, 2 (2015).

549. Zhao, G. *et al.* VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**, 21–30 (2017).

550. Bzhalava, D., Eklund, C. & Dillner, J. International standardization and classification of human papillomavirus types. *Virology* **476**, 341–344 (2015).

551. Smelov, V. *et al.* Prevalence of cutaneous beta and gamma human papillomaviruses in the anal canal of men who have sex with women. *Papillomavirus Research* **3**, 66–72 (2017).

552. Bernard, H.-U. *et al.* Classification of Papillomaviruses (PVs) Based on 189 PV Types and Proposal of Taxonomic Amendments. *Virology* **401**, 70–79 (2010).

553. Haedicke, J. & Iftner, T. Human papillomaviruses and cancer. *Radiotherapy and Oncology* **108**, 397–402 (2013).

554. Bouvard, V. *et al.* A review of human carcinogens--Part B: biological agents. *Lancet Oncol.* **10**, 321–322 (2009).

555. Giuliano, A. R. *et al.* Epidemiology of Human Papillomavirus Infection in Men, in Cancers other than Cervical and in Benign Conditions. *Vaccine* **26**, K17–K28 (2008).

556. Goon, P., Sonnex, C., Jani, P., Stanley, M. & Sudhoff, H. Recurrent respiratory papillomatosis: an overview of current thinking and treatment. *Eur Arch Otorhinolaryngol* **265**, 147–151 (2008).

557. Pfister, H. Chapter 8: Human papillomavirus and skin cancer. *J. Natl. Cancer Inst. Monographs* 52–56 (2003).

558. Andersson, K. *et al.* Seroreactivity to cutaneous human papillomaviruses among patients with nonmelanoma skin cancer or benign skin lesions. *Cancer Epidemiol. Biomarkers Prev.* **17**, 189–195 (2008).

559. Berkhout, R. J., Bouwes Bavinck, J. N. & ter Schegget, J. Persistence of human papillomavirus DNA in benign and (pre)malignant skin lesions from renal transplant recipients. *J. Clin. Microbiol.* **38**, 2087–2096 (2000).

560. Bouwes Bavinck, J. N. *et al.* Multicenter study of the association between betapapillomavirus infection and cutaneous squamous cell carcinoma. *Cancer Res.* **70**, 9777–9786 (2010).

561. Casabonne, D. *et al.* A prospective pilot study of antibodies against human papillomaviruses and cutaneous squamous cell carcinoma nested in the Oxford component of the European Prospective Investigation into Cancer and Nutrition. *Int. J. Cancer* **121**, 1862–1868 (2007).

562. de Jong-Tieben, L. M. *et al.* High frequency of detection of epidermodysplasia verruciformis-associated human papillomavirus DNA in biopsies from malignant and premalignant skin lesions from renal transplant recipients. *J. Invest. Dermatol.* **105**, 367–371 (1995).

563. Harwood, C. A. *et al.* Human papillomavirus infection and non-melanoma skin cancer in immunosuppressed and immunocompetent individuals. *J. Med. Virol.* **61**, 289–297 (2000).

564. Iftner, A. *et al.* The prevalence of human papillomavirus genotypes in nonmelanoma skin cancers of nonimmunosuppressed individuals identifies high-risk genital types as possible risk factors. *Cancer Res.* **63**, 7515–7519 (2003).

565. Karagas, M. R. *et al.* Human papillomavirus infection and incidence of squamous cell and basal cell carcinomas of the skin. *J. Natl. Cancer Inst.* **98**, 389–395 (2006).

566. Waterboer, T. *et al.* Serological association of beta and gamma human papillomaviruses with squamous cell carcinoma of the skin. *Br. J. Dermatol.* **159**, 457–459 (2008).

567. Cornet, I. *et al.* Comparative analysis of transforming properties of E6 and E7 from different beta human papillomavirus types. *J. Virol.* **86**, 2366–2370 (2012).

568. Pierce Campbell, C. M. *et al.* Cutaneous beta human papillomaviruses and the development of male external genital lesions: A case-control study nested within the HIM Study. *Virology* **497**, 314–322 (2016).

569. Donà, M. G. *et al.* Incidence, clearance and duration of cutaneous beta and gamma human papillomavirus anal infection. *J. Infect.* **73**, 380–383 (2016).

570. Hampras, S. S. *et al.* Prevalence and Concordance of Cutaneous Beta Human Papillomavirus Infection at Mucosal and Cutaneous Sites. *J. Infect. Dis.* **216**, 92–96 (2017).

571. Forslund, O., Johansson, H., Madsen, K. G. & Kofoed, K. The Nasal Mucosa Contains a Large Spectrum of Human Papillomavirus Types from the Betapapillomavirus and Gammapapillomavirus Genera. *J Infect Dis* **208**, 1335–1341 (2013).

572. Viarisio, D. *et al.* Novel ß-HPV49 Transgenic Mouse Model of Upper Digestive Tract Cancer. *Cancer Res.* **76**, 4216–4225 (2016).

573. de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. & zur Hausen, H. Classification of papillomaviruses. *Virology* **324**, 17–27 (2004).

574. Saito, J. *et al.* New human papillomavirus sequences in female genital tumors from Japanese patients. *Jpn. J. Cancer Res.* **78**, 1081–1087 (1987).

575. Beaudenon, S. *et al.* A new type of human papillomavirus associated with oral focal epithelial hyperplasia. *J. Invest. Dermatol.* **88**, 130–135 (1987).

576. Lorincz, A. T., Lancaster, W. D. & Temple, G. F. Cloning and characterization of the DNA of a new human papillomavirus from a woman with dysplasia of the uterine cervix. *Journal of Virology* **58**, 225–229 (1986).

577. Kahn, T., Schwarz, E. & zur Hausen, H. Molecular cloning and characterization of the DNA of a new human papillomavirus (HPV 30) from a laryngeal carcinoma. *Int. J. Cancer* **37**, 61–65 (1986).

578. Favre, M. *et al.* Two new human papillomavirus types (HPV54 and 55) characterized from genital tumours illustrate the plurality of genital HPVs. *Int. J. Cancer* **45**, 40–46 (1990).

579. Grimmel, M., de Villiers, E. M., Neumann, C., Pawlita, M. & zur Hausen, H. Characterization of a new human papillomavirus (HPV 41) from disseminated warts and detection of its DNA in some skin carcinomas. *Int. J. Cancer* **41**, 5–9 (1988).

580. Orth, G., Favre, M. & Croissant, O. Characterization of a new type of human papillomavirus that causes skin warts. *J. Virol.* **24**, 108–120 (1977).

581. Illumina. An introduction to Next-Generation Sequencing Technology. (2017).

582. Dutta, S., Robitaille, A., Rollison, D. E., Tommasino, M. & Gheit, T. Complete Genome Sequence of a Novel Human Betapapillomavirus Isolated from a Skin Sample. *Genome Announc* **5**, (2017).

583. Hampras, S. S. *et al.* Natural history of polyomaviruses in men: the HPV infection in men (HIM) study. *J. Infect. Dis.* **211**, 1437–1446 (2015).

584. Hampras, S. S. *et al.* Natural History of Cutaneous Human Papillomavirus (HPV) Infection in Men: The HIM Study. *PLoS One* **9**, (2014).

585. Nunes, E. M. *et al.* Diversity of beta-papillomavirus at anogenital and oral anatomic sites of men: The HIM Study. *Virology* **495**, 33–41 (2016).

586. Pierce Campbell, C. M. *et al.* Cutaneous human papillomavirus types detected on the surface of male external genital lesions: a case series within the HPV Infection in Men Study. *J. Clin. Virol.* **58**, 652–659 (2013).

587. Giuliano, A. R. *et al.* Incidence and clearance of genital human papillomavirus infection in men (HIM): a cohort study. *Lancet* **377**, 932–940 (2011).

588. Giuliano, A. R. *et al.* The Human Papillomavirus Infection in Men Study: Human Papillomavirus Prevalence and Type Distribution among Men Residing in Brazil, Mexico, and the United States. *Cancer Epidemiol Biomarkers Prev* **17**, 2036–2043 (2008).

589. Giuliano, A. *et al.* Circumcision and Sexual Behavior: Factors Independently Associated with Human Papillomavirus (HPV) Detection among Men in The HIM Study. *Int J Cancer* **124**, 1251–1257 (2009).

590. Nyitray, A. G. *et al.* Age-Specific Prevalence of and Risk Factors for Anal Human Papillomavirus (HPV) among Men Who Have Sex with Women and Men Who Have Sex with Men: The HPV in Men (HIM) Study. *J Infect Dis* **203**, 49–57 (2011).

591. Chouhy, D. *et al.* New generic primer system targeting mucosal/genital and cutaneous human papillomaviruses leads to the characterization of HPV 115, a novel Beta-papillomavirus species 3. *Virology* **397**, 205–216 (2010).

592. Forslund, O., Antonsson, A., Nordin, P., Stenquist, B. & Hansson, B. G. A broad range of human papillomavirus types detected with a general PCR method suitable for analysis of cutaneous tumours and normal skin. *J. Gen. Virol.* **80 (Pt 9)**, 2437–2443 (1999).

593. Käller, M, . Ewels, P. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016).

594. Mahé, F., . Rognes, T. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* (2016).

595. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).

596. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

597. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

598. Nei and Kumar. *Molecular evolution and phylogenetics*. (Oxford University Press, 2000).

599. Berger, S. A. & Stamatakis, A. Aligning short reads to reference alignments and trees. *Bioinformatics* **27**, 2068–2075 (2011).

600. Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* **60**, 291–302 (2011).

601. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

602. Matsen, F. A., Hoffman, N. G., Gallagher, A. & Stamatakis, A. A format for phylogenetic placements. *PLoS ONE* **7**, e31009 (2012).

603. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

604. Madden, T. *The BLAST Sequence Analysis Tool*. (National Center for Biotechnology Information (US), 2003).

605. Berger,S.A. & Stamatakis,A. PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension.

606. Schowalter, R. M., Pastrana, D. V., Pumphrey, K. A., Moyer, A. L. & Buck, C. B. Merkel Cell Polyomavirus and Two Novel Polyomaviruses Are Chronically Shed from Human Skin. *Cell Host Microbe* **7**, 509–515 (2010).

607. Rector, A., Tachezy, R. & Van Ranst, M. A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. *J. Virol.* **78**, 4993–4998 (2004).

608. Batovska, J., Lynch, S. E., Rodoni, B. C., Sawbridge, T. I. & Cogan, N. O. Metagenomic arbovirus detection using MinION nanopore sequencing. *J. Virol. Methods* **249**, 79–84 (2017).

609. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).

610. Adrien Leger and Tommaso Leonardi. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *JOSS* (2019).

611. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

612. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

613. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).

614. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

615. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

616. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

617. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

618. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

619. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–3500 (2003).

620. Myers, G., G. M., C. Baker, K. Münger, F. Sverdrup, A. McBride &. H. U. Bernard. Alignments. In Human Papillomaviruses 1997. HPV Sequence Database. II-L1–23–73 (1997).

621. Myers, G., G. M., C. Baker, K. Münger, F. Sverdrup, A. McBride &. H. U. Bernard. Alignments. In Human Papillomaviruses 1996. HPV Sequence Database. II-L1–1–67 (1996).

622. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

623. Dutta, S. *et al.* Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. *Genome Announc* **5**, (2017).

624. Brancaccio, R. N. *et al.* Complete Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. *Genome Announc* **5**, (2017).

625. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).

626. Greninger, A. L. *et al.* Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* **7**, 99 (2015).

627. Tyler, A. D. *et al.* Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci Rep* **8**, 10931 (2018).

628. Orth, G. *et al.* Characterization of two types of human papillomaviruses in lesions of epidermodysplasia verruciformis. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 1537–1541 (1978).

629. Kocjan, B. J., Bzhalava, D., Forslund, O., Dillner, J. & Poljak, M. Molecular methods for identification and characterization of novel papillomaviruses. *Clin. Microbiol. Infect.* **21**, 808–816 (2015).

630. Arroyo, L. S. *et al.* Next generation sequencing for human papillomavirus genotyping. *J. Clin. Virol.* **58**, 437–442 (2013).

631. Barzon, L. *et al.* Human papillomavirus genotyping by 454 next generation sequencing technology. *J. Clin. Virol.* **52**, 93–97 (2011).

632. Johansson, H. *et al.* Metagenomic sequencing of 'HPV-negative' condylomas detects novel putative HPV types. *Virology* **440**, 1–7 (2013).

633. Pfister, H. Chapter 8: Human papillomavirus and skin cancer. *J. Natl. Cancer Inst. Monographs* 52–56 (2003).

634. Bravo, I. G. & Félez-Sánchez, M. Papillomaviruses: Viral evolution, cancer and evolutionary medicine. *Evol Med Public Health* **2015**, 32–51 (2015).

635. Gottschling, M. *et al.* Quantifying the phylodynamic forces driving papillomavirus evolution. *Mol. Biol. Evol.* **28**, 2101–2113 (2011).

636. Lunardi, M. *et al.* Bovine papillomavirus type 13 DNA in equine sarcoids. *J. Clin. Microbiol.* **51**, 2167–2171 (2013).

637. Trewby, H. *et al.* Analysis of the long control region of bovine papillomavirus type 1 associated with sarcoids in equine hosts indicates multiple cross-species transmission events and phylogeographical structure. *J. Gen. Virol.* **95**, 2748–2756 (2014).

638. García-Pérez, R. *et al.* Novel papillomaviruses in free-ranging Iberian bats: no virus-host co-evolution, no strict host specificity, and hints for recombination. *Genome Biol Evol* **6**, 94–104 (2014).

639. Chen, Z. *et al.* Genomic diversity and interspecies host infection of alpha12 Macaca fascicularis papillomaviruses (MfPVs). *Virology* **393**, 304–310 (2009).

640. Anis, E. A. *et al.* Molecular characterization of the L1 gene of papillomaviruses in epithelial lesions of cats and comparative analysis with corresponding gene sequences of human and feline papillomaviruses. *Am. J. Vet. Res.* **71**, 1457–1461 (2010).

641. O'Neill, S. H. *et al.* Detection of human papillomavirus DNA in feline premalignant and invasive squamous cell carcinoma. *Vet. Dermatol.* **22**, 68–74 (2011).

642. Rector, A. & Van Ranst, M. Animal papillomaviruses. *Virology* **445**, 213–223 (2013).

643. Vanmechelen, B. *et al.* Genomic characterization of Erethizon dorsatum papillomavirus 2, a new papillomavirus species marked by its exceptional genome size. *J. Gen. Virol.* **99**, 1699–1704 (2018).

644. Schmitt, M. *et al.* Multiple human papillomavirus infections with high viral loads are associated with cervical lesions but do not differentiate grades of cervical abnormalities. *J. Clin. Microbiol.* **51**, 1458–1464 (2013).

645. Arroyo, L. S. *et al.* Next generation sequencing for human papillomavirus genotyping. *J. Clin. Virol.* **58**, 437–442 (2013).

646. Barzon, L. *et al.* Human papillomavirus genotyping by 454 next generation sequencing technology. *J. Clin. Virol.* **52**, 93–97 (2011).

647. Brancaccio, R. N. *et al.* Generation of a novel next-generation sequencing-based method for the isolation of new human papillomavirus types. *Virology* **520**, 1–10 (2018).

648. Bzhalava, D. *et al.* Deep sequencing extends the diversity of human papillomaviruses in human skin. *Sci Rep* **4**, 5807 (2014).

649. Kocjan, B. J., Bzhalava, D., Forslund, O., Dillner, J. & Poljak, M. Molecular methods for identification and characterization of novel papillomaviruses. *Clin. Microbiol. Infect.* **21**, 808–816 (2015).

650. Ekström, J., Bzhalava, D., Svenback, D., Forslund, O. & Dillner, J. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. *Int. J. Cancer* **129**, 2643–2650 (2011).

651. Johansson, H. *et al.* Metagenomic sequencing of 'HPV-negative' condylomas detects novel putative HPV types. *Virology* **440**, 1–7 (2013).

652. McNaughton, A. L. *et al.* Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Sci Rep* **9**, 7081 (2019).

653. John Beaulaurier, Edward F. DeLong, (last) & Elaine Luo, John Eppley, Paul Den Uyl, Xiaoguang Dai, Daniel J Turner, Matthew Pendelton, Sissel Juul, Eoghan Harrington. Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities. *BioRxiv* (2019).

654. Kuiama Lewandowski, Philippa C. Matthews, (last), Yifei Xu, Steven .T Pullan, Sheila F. Lumley, Dona Foster, Nicholas Sanderson, Alison Vaughan, Marcus Morgan, Nicole Bright, James Kavanagh, Richard Vipond, Miles Carroll, Anthony C. Marriott, Karen E Gooch, Monique Andersson, Katie Jeffery, Timothy EA Peto & Derrick W. Crook, A Sarah Walker.

Metagenomic Nanopore sequencing of influenza virus direct from clinical respiratory samples. *BioRxiv* (2019).

655. Pastrana, D. V. *et al.* Metagenomic Discovery of 83 New Human Papillomavirus Types in Patients with Immunodeficiency. *mSphere* **3**, (2018).
656. 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* **24**, 335–341 (2018).
657. Augier, C. *et al.* Identification of a Novel Simian Immunodeficiency Virus-Infected African Green Monkey (Chlorocebus tantalus) Confirms that Tantalus Monkeys in Cameroon Are Infected with a Mosaic SIVagm Lineage. *AIDS Res. Hum. Retroviruses* (2019).
658. Wawina-Bokalanga, T. *et al.* Complete Genome Sequence of a New Ebola Virus Strain Isolated during the 2017 Likati Outbreak in the Democratic Republic of the Congo. *Microbiol Resour Announc* **8**, (2019).
659. Uchida, Y. *et al.* A case of genotype-3b hepatitis C virus in which the whole genome was successfully analyzed using third-generation nanopore sequencing. *Hepatol. Res.* **49**, 1083–1087 (2019).
660. Karamitros, T. *et al.* De Novo Assembly of Human Herpes Virus Type 1 (HHV-1) Genome, Mining of Non-Canonical Structures and Detection of Novel Drug-Resistance Mutations Using Short- and Long-Read Next Generation Sequencing Technologies. *PLoS ONE* **11**, e0157600 (2016).
661. Prazsák, I. *et al.* Full Genome Sequence of the Western Reserve Strain of Vaccinia Virus Determined by Third-Generation Sequencing. *Genome Announc* **6**, (2018).
662. Günther, T. *et al.* Recovery of the first full-length genome sequence of a parapoxvirus directly from a clinical sample. *Sci Rep* **7**, 3734 (2017).
663. de Jesus, J. G., Giovanetti, M., Rodrigues Faria, N. & Alcantara, L. C. J. Acute Vector-Borne Viral Infection: Zika and MinION Surveillance. *Microbiol Spectr* **7**, (2019).

664. Vanmechelen, B. *et al.* Identification of a novel species of papillomavirus in giraffe lesions using nanopore sequencing. *Vet. Microbiol.* **201**, 26–31 (2017).

665. Boldogkői, Z. *et al.* Transcriptomic study of Herpes simplex virus type-1 using full-length sequencing techniques. *Sci Data* **5**, 180266 (2018).

666. Prazsák, I. *et al.* Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* **19**, 873 (2018).

667. Tombácz, D. *et al.* Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Front Genet* **10**, 834 (2019).

668. Tombácz, D., Balázs, Z., Csabai, Z., Snyder, M. & Boldogkői, Z. Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses. *Front Genet* **9**, 259 (2018).

669. Tombácz, D. *et al.* Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* **7**, (2018).

670. Boldogkői, Z., Moldován, N., Szűcs, A. & Tombácz, D. Transcriptome-wide analysis of a baculovirus using nanopore sequencing. *Sci Data* **5**, 180276 (2018).

671. Moldován, N. *et al.* Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *Sci Rep* **8**, 8604 (2018).

672. Croville, G. *et al.* Rapid whole-genome based typing and surveillance of avipoxviruses using nanopore sequencing. *J. Virol. Methods* **261**, 34–39 (2018).

673. Yinda, C. K. *et al.* A Novel Field-Deployable Method for Sequencing and Analyses of Henipavirus Genomes From Complex Samples on the MinION Platform. *J. Infect. Dis.* (2019).

674. Sasani, T. A., Cone, K. R., Quinlan, A. R. & Elde, N. C. Long read sequencing reveals poxvirus evolution through rapid homogenization of gene arrays. *Elife* **7**, (2018).

675. Bainomugisa, A. *et al.* A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genom* **4**, (2018).

676. Keller, M. W. *et al.* Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Sci Rep* **8**, 14408 (2018).
677. Sutton, K. M. *et al.* Detection of atypical porcine pestivirus genome in newborn piglets affected by congenital tremor and high preweaning mortality1. *J. Anim. Sci.* **97**, 4093–4100 (2019).
678. Brancaccio, R. N. *et al.* Isolation of a Novel Beta-2 Human Papillomavirus from Skin. *Microbiol Resour Announc* **8**, (2019).
679. Brancaccio, R. N. *et al.* Complete Genome Sequence of a Novel Human Gammapapillomavirus Isolated from Skin. *Genome Announc* **5**, (2017).

Supplementary data

Table S1: PV-related sequences in skin samples. Each read was identified using the Evolutionary Placement Algorithm in RAxML.

Specimen	Classification			Number of reads (N)				Pool
	PV Genus	PV Species	PV Related	Beta-3-1/Beta-3-2	FAP	FAPM1	CUT	
Skin	Alphapapillomavirus	Alphapapillomavirus 11	HPV73	0	0	441	0	3
		Alphapapillomavirus 14	CgPV1	1465375	0	0	0	1
		Alphapapillomavirus 2	HPV28	0	0	2156	92225	3,4
			HPV29	0	0	0	79	4
			HPV3	0	0	196927	0	3
			HPV77	0	0	0	14652	4
		Alphapapillomavirus 3	HPV114	0	0	0	7223	4
			HPV83	0	0	0	2	4
			HPV84	0	0	121	0	3
			HPV89	0	0	7	0	3
		Alphapapillomavirus 4	HPV2	0	0	0	2	4
			HPV27	0	0	0	8	4
			HPV57	0	0	0	2	4
		Alphapapillomavirus 5	HPV51	0	0	0	10	4
			HPV82	0	254	0	0	2
		Alphapapillomavirus 6	HPV30	0	0	0	8	4
			HPV66	0	82	0	219	2,4
		Alphapapillomavirus 7	HPV45	0	0	7	0	3
			HPV68	0	0	1232	0	3
		Alphapapillomavirus 8	HPV40	0	0	2	0	3
			HPV7	0	24	0	0	2
		Alphapapillomavirus 9	HPV16	2	0	0	0	1
	Betapapillomavirus	Betapapillomavirus	HPV-mEV03c09	0	0	2	145	3,4
			HPV-mHIVGc36	0	0	22	0	3
			HPV-mMTS1	0	368	0	0	2
			HPV-mm292c10	0	0	34	571	3,4
			HPV-mm292c100	0	0	12	12	3,4
			HPV-mm292c14	0	0	2652	0	3
			HPV-mm292c88	0	7826	480	0	2,3
			HPV-mw15c111	0	0	0	37828	4
			HPV209	0	0	0	116	4
		Betapapillomavirus 1	HPV105	0	27706	2137	892	2,3,4
			HPV118	0	17	4	222	2,3,4
			HPV12	0	90	6447	8545	2,3,4
			HPV124	0	2	7611	0	2,3
			HPV14	0	4104	0	953	2,4
			HPV143	0	16	0	0	2
			HPV152	0	10	2	1592	2,3,4
			HPV19	0	8814	0	327	2,4
			HPV20	0	32	381	104	2,3,4
			HPV21	0	59	205	45002	2,3,4
			HPV24	0	27356	3658	575	2,3,4
			HPV25	0	60	0	0	2
			HPV36	0	7457	11249	157	2,3,4
			HPV47	0	6	0	11395	2,4
			HPV5	0	62	710	123	2,3,4
			HPV8	0	220	10840	8278	2,3,4
			HPV93	0	0	1835	2	3,4

			HPV98	0	20	542	0	2,3
		Betapapillomavirus 2	HPV100	0	6	1811	0	2,3
			HPV104	0	12	521	45	2,3,4
			HPV107	0	18420	0	19	2,4
			HPV110	0	2	510	0	2,3
			HPV111	0	33425	3827	0	2,3
			HPV120	0	205139	24089	197	2,3,4
			HPV122	0	0	1153	12	3,4
			HPV145	0	2	23	15225	2,3,4
			HPV15	0	0	93	14338	3,4
			HPV151	0	0	119	0	3
			HPV17	0	157	279	6039	2,3,4
			HPV174	573	0	47	55	1,3,4
			HPV22	0	0	0	2	4
			HPV23	0	88438	0	1341	2,4
			HPV37	0	38664	324	114	2,3,4
			HPV38	0	459	4255	0	2,3
			HPV9	0	50	78	3877	2,3,4
		Betapapillomavirus 3	HPV115	323	0	0	25	1,4
			HPV49	796696	0	30739	6	1,3,4
			HPV75	0	9	0	0	2
			HPV76	128	5580	196	0	1,2,3
		Betapapillomavirus 4	HPV92	0	0	36	3043	3,4
		Betapapillomavirus 5	HPV150	0	0	0	12	4
			HPV96	80	6842	845	11739	1,2,3,4
	Chipapillomavirus	Chipapillomavirus 1	CPV5	0	4	0	653	2,4
			CPV9	0	726	17	0	2,3
		Chipapillomavirus 2	CPV16	0	0	32	0	3
			CPV4	0	750	0	0	2
		Chipapillomavirus 3	CPV10	0	0	62	0	3
			CPV14	0	0	0	13	4
			CPV8	0	75	935	30	2,3,4
	Deltapapillomavirus	Deltapapillomavirus 1	AaPV1	0	0	52	0	3
		Deltapapillomavirus 2	OvPV1	18	3555	50160	43571	1,2,3,4
		Deltapapillomavirus 4	BPV14	0	0	246	0	3
			BPV2	0	6	0	0	2
		Deltapapillomavirus 5	CcaPV1	0	4	0	0	2
		Deltapapillomavirus 6	CdPV1	0	0	15508	0	3
			CdPV2	0	2	0	0	2
	Dyoepsilonpapillomavirus	Dyoepsilonpapillomavirus 1	FlPV1	0	36510	472	1264	2,3,4
	Dyoiotapapillomavirus	Dyoiotapapillomavirus 2	EcPV4	0	0	8	0	3
			EcPV5	0	0	0	14	4
	Dyokappapapillomavirus	Dyokappapapillomavirus	BPV16	0	893	1461	2	2,3,4
			BPV18	0	0	2047	0	3
			BPV22	0	0	10	0	3
		Dyokappapapillomavirus 1	OaPV3	0	295	0	1328	2,4
	Dyonupapillomavirus	Dyonupapillomavirus 1	ZcPV1	0	26	0	0	2
	Dyoomegapapillomavirus	Dyoomegapapillomavirus 1	EsPV2	0	0	0	514	4

	Dyoomikronpapillomavirus	Dyoomikronpapillomavirus 1	SscPV1	0	0	639	0	3
			SscPV3	26265	0	0	114	1,4
	Dyophipapillomavirus	Dyophipapillomavirus 1	TePV1	6	0	0	0	1
	Dyopipapillomavirus	Dyopipapillomavirus 1	PphPV4	0	0	138	0	3
	Dyorhopapillomavirus	Dyorhopapillomavirus 1	EcPV3	0	0	0	4	4
			EcPV6	0	0	0	25	4
			EcPV7	0	3	0	0	2
	Dyosigmapapillomavirus	Dyosigmapapillomavirus 1	CcanPV1	0	0	0	8	4
	Dyoupsilonpapillomavirus	Dyoupsilonpapillomavirus 1	EhPV1	0	0	25345	26280	3,4
	Dyoxipapillomavirus	Dyoxipapillomavirus 1	BPV7	0	0	0	1130	4
	Dyozetapapillomavirus	Dyozetapapillomavirus 1	CcPV1	0	20	10	53455	2,3,4
			CmPV1	0	79	6	0	2,3
	Epsilonpapillomavirus	Epsilonpapillomavirus 1	BPV5	0	0	80	0	3
	Etapapillomavirus	Etapapillomavirus 1	FcPV1	0	0	204	0	3
	Gammapapillomavirus	Gammapapillomavirus	HPV-mCG3	0	0	0	36388	4
			HPV-mCH2	0	0	0	135	4
			HPV-mDysk1	644	0	0	0	1
			HPV-mDysk2	0	0	0	1964	4
			HPV-mDysk3	0	139	0	4677	2,4
			HPV-mDysk5	0	45	0	0	2
			HPV-mDysk6	2	0	261	0	1,3
			HPV-mEV03c104	0	0	6	0	3
			HPV-mEV03c212	0	36685	0	0	2
			HPV-mEV03c40	0	1477	68	1298	2,3,4
			HPV-mEV03c434	0	6790	11964	46171	2,3,4
			HPV-mEV03c60	0	0	156	0	3
			HPV-mEV07c367	0	0	0	16442	4
			HPV-mEV07c382	0	2	0	0	2
			HPV-mEV07c390	0	13593	0	368	2,4
			HPV-mFD1	0	7390	0	0	2
			HPV-mFD2	0	0	81	155	3,4
			HPV-mFS1	0	73	108	0	2,3
			HPV-mHIVGc70	0	0	0	40	4
			HPV-mICB1	0	0	2	0	3
			HPV-mKC5	0	0	418	2	3,4
			HPV-mKN1	0	0	8	457	3,4
			HPV-mL55	0	0	6	0	3
			HPV-mMTS2	0	0	8	2	3,4
			HPV-mSE355	0	0	66	0	3
			HPV-mSE379	0	39770	0	4	2,4
			HPV-mSE383	0	1573	0	0	2
			HPV-mTVMBSGc2450	0	0	2896	0	3
			HPV-mTVMBSGc529	0	43	12	3280	2,3,4
			HPV-mTVMBSHc13	0	1286	0	1771	2,4
			HPV-mTVMBSHc33	0	16	0	0	2
			HPV-mTVMBSWc141	0	1340	41	0	2,3
			HPV-mdo1c232	0	0	0	60	4
			HPV-mga2c01	0	0	0	76	4

			HPV-mm090c10	0	0	0	8	4
			HPV-mm090c145	0	0	0	2	4
			HPV-mm090c66	0	0	29	0	3
			HPV-mw03c65	0	0	0	93	4
			HPV-mw07c34d	0	0	0	38	4
			HPV-mw11C39	0	82	289	1011	2,3,4
			HPV-mw11C51	0	0	0	10	4
			HPV-mw18c11d	0	15	0	12	2,4
			HPV-mw18c25	0	0	2	0	3
			HPV-mw18c39	0	0	0	18	4
			HPV-mw20c01b	0	0	2714	0	3
			HPV-mw20c02c	0	27	10	16	2,3,4
			HPV-mw20c04	0	0	82	12	3,4
			HPV-mw20c08a	0	0	0	8432	4
			HPV-mw21c693	0	0	17	0	3
			HPV-mw23c08c	0	0	0	8	4
			HPV-mw23c77	0	0	0	4	4
			HPV-mw27c04c	0	2	0	0	2
			HPV-mw27c157c	0	8	0	0	2
			HPV-mw27c39c	0	121	2	6653	2,3,4
			HPV-mw34c04a	0	3	0	0	2
			HPV-mw34c11a	0	0	10	17421	3,4
			HPV-mw34c28a	0	0	163	0	3
			HPV-mw34c34a	0	27	2	1778	2,3,4
			HPV-mwg1c05	0	210	1872	39034	2,3,4
			HPV-mwg1c09	0	22	36	0	2,3
			HPV157	0	0	0	612	4
			HPV205	0	0	0	4	4
		Gammapapillomavirus 1	HPV173	0	0	0	210	4
			HPV95	0	205	0	0	2
		Gammapapillomavirus 10	HPV130	0	1164	45	0	2,3
			HPV142	0	6	0	24	2,4
			HPV180	0	0	254	0	3
		Gammapapillomavirus 11	HPV126	0	0	178	0	3
			HPV136	0	148	417	26	2,3,4
			HPV140	0	16	0	0	2
			HPV141	0	0	78	8	3,4
			HPV154	0	0	0	686	4
			HPV171	0	2	0	0	2
			HPV202	0	0	2	0	3
		Gammapapillomavirus 12	HPV127	0	0	0	46	4
			HPV132	0	0	45	0	3
			HPV148	0	2135	4505	2530	2,3,4
			HPV165	0	16267	982	2550	2,3,4
			HPV199	0	0	3	0	3
		Gammapapillomavirus 13	HPV128	0	2	0	0	2
			HPV153	0	0	0	36	4
		Gammapapillomavirus 14	HPV131	0	0	0	2	4

		Gammapapillomavirus 15	HPV135	0	99	110	10	2,3,4
			HPV146	0	0	141	0	3
			HPV179	0	0	0	55	4
		Gammapapillomavirus 16	HPV137	0	2	75711	0	2,3
		Gammapapillomavirus 17	HPV144	0	75	0	339	2,4
		Gammapapillomavirus 19	HPV162	0	6	0	0	2
			HPV166	0	0	2	0	3
		Gammapapillomavirus 2	HPV200	0	192	0	35	2,4
		Gammapapillomavirus 21	HPV167	0	48	53	1488	2,3,4
		Gammapapillomavirus 22	HPV172	0	0	2	0	3
			HPVMTS4	0	0	258	0	3
		Gammapapillomavirus 23	HPV175	0	0	3107	2	3,4
		Gammapapillomavirus 24	HPV178	0	0	0	2	4
			HPV197	0	110	0	77	2,4
		Gammapapillomavirus 25	HPV184	0	11475	12792	248	2,3,4
		Gammapapillomavirus 27	HPV201	0	6	4	1162	2,3,4
		Gammapapillomavirus 3	HPV50	0	2	1501	3356	2,3,4
		Gammapapillomavirus 4	HPV60	0	16	0	0	2
		Gammapapillomavirus 6	HPV101	0	0	2	0	3
		Gammapapillomavirus 7	HPV109	0	0	111	5009	3,4
			HPV123	0	66	2	1341	2,3,4
			HPV134	0	2	11	0	2,3
			HPV138	0	2714	0	0	2
			HPV139	0	0	1484	6	3,4
			HPV149	0	0	0	445	4
			HPV155	0	2	0	0	2
		Gammapapillomavirus 8	HPV112	0	0	5	66	3,4
			HPV119	0	0	176	2	3,4
			HPV147	0	0	76	0	3
			HPV164	0	0	0	155	4
			HPV168	0	0	0	4	4
		Gammapapillomavirus 9	HPV116	0	0	12	4	3,4
			HPV129	0	0	47	21	3,4
	Iotapapillomavirus	Iotapapillomavirus 1	RnPV3	0	0	0	31	4
	Kappapapillomavirus	Kappapapillomavirus 2	SfPV1	0	8	6844	0	2,3
	Lambdapapillomavirus	Lambdapapillomavirus	AmPV4	0	859	0	0	2
		Lambdapapillomavirus 2	CPV1	0	549	8566	0	2,3
		Lambdapapillomavirus 3	CPV6	0	0	7	2	3,4
		Lambdapapillomavirus 5	CcrPV1	0	0	0	37538	4
	Mupapillomavirus	Mupapillomavirus	HPV-md01c06	0	0	0	2	4
		Mupapillomavirus 1	HPV1	0	311	18250	6	2,3,4
		Mupapillomavirus 2	HPV63	0	0	47403	0	3
	Nupapillomavirus	Nupapillomavirus 1	HPV41	0	0	0	12	4
	Omegapapillomavirus	Omegapapillomavirus	AmPV1	0	0	0	2	4
	Omikronpapillomavirus	Omikronpapillomavirus 1	PsPV1	0	2	0	0	2
	Pipapillomavirus	Pipapillomavirus 1	MaPV1	0	0	24	0	3
		Pipapillomavirus 2	MmuPV1	0	0	20	0	3
			RnPV1	0	284	0	0	2

Psipapillomavirus	Psipapillomavirus	EhPV2	0	10781	2	8556	2,3,4
	Psipapillomavirus 1	RaPV1	0	0	0	968	4
Rhopapillomavirus	Rhopapillomavirus 1	TmPV2	0	5	0	137	2,4
Sigmapapillomavirus	Sigmapapillomavirus 1	EdPV1	0	1895	26153	70	2,3,4
Taupapillomavirus	Taupapillomavirus 1	CPV2	0	82	0	0	2
	Taupapillomavirus 2	CPV13	0	1207	0	0	2
	Taupapillomavirus 3	FcaPV4	0	111233	6	0	2,3
Thetapapillomavirus	Thetapapillomavirus 1	PePV1	0	150	6058	674	2,3,4
Treisdeltapapillomavirus	Treisdeltapapillomavirus 1	RfPV1	0	0	2	0	3
Treisepsilonpapillomavirus	Treisepsilonpapillomavirus	PaPV2	0	21	12100	609	2,3,4
	Treisepsilonpapillomavirus 1	PaPV1	111484	157	2531	3006	1,2,3,4
Treisetapapillomavirus	Treisetapapillomavirus	VvPV1	0	0	0	8	4
Treiszetapapillomavirus	Treiszetapapillomavirus	FgPV1	0	91027	1006	2043	2,3,4
Unclassified	Unclassified	BPV19	0	0	5	0	3
		EcPV8	0	0	15364	889	3,4
		MrPV1	0	0	3863	0	3
		PpuPV1	0	0	6	0	3
		SaPV1	187135	98743	172192	208682	1,2,3,4
Upsilonpapillomavirus	Upsilonpapillomavirus 1	TtPV3	0	12	180	404	2,3,4
		TtPV4	0	0	0	12	4
	Upsilonpapillomavirus 3	PphPV2	0	78	0	0	2
Xipapillomavirus	Xipapillomavirus	BPV20	0	605	0	4	2,4
	Xipapillomavirus 1	BPV15	0	16	2109	23	2,3,4
		BPV3	0	630	0	0	2
Zetapapillomavirus	Zetapapillomavirus 1	EcPV1	0	0	0	6	4

Table S2: PV-related sequences in oral samples. Each read was identified using the Evolutionary Placement Algorithm in RAxML

Specimen	Classification			Number of reads (N)				Pool
	PV Genus	PV Species	PV Related	FAP M1	FAP M2	CUT	MIX*	
Oral	Alphapapillomavirus	Alphapapillomavirus 1	HPV42	0	0	0	147488	8
		Alphapapillomavirus 10	HPV6	0	0	0	13	8
		Alphapapillomavirus 2	HPV10	0	0	968	0	7
			HPV3	0	272	0	0	5
			HPV77	0	0	2	0	7
			HPV94	0	184	0	0	5
		Alphapapillomavirus 3	HPV84	0	0	665	0	7
		Alphapapillomavirus 4	HPV57	0	0	8	0	7
		Alphapapillomavirus 5	HPV26	8	0	0	11366	6,8
			HPV51	0	0	2	0	7
			HPV69	0	0	0	6648	8
		Alphapapillomavirus 6	HPV56	0	0	0	82	8
			HPV66	99	1131	2	387	5,6,7,8
		Alphapapillomavirus 7	HPV45	18	18126	0	0	5,6
		Alphapapillomavirus 9	HPV16	98	23645	0	2	5,6,8
			HPV52	0	0	0	2	8
	Betapapillomavirus	Betapapillomavirus	HPV-mHIVGc36	307	28775	712	0	5,6,7
			HPV-mMTS1	129	53	0	260	5,6,8
			HPV-mRTRX7	1542	0	0	0	6
			HPV-mm292c10	0	0	98	73	7,8
			HPV-mm292c14	0	0	0	29294	8
			HPV-mw15c111	2	0	0	2	6,8
		Betapapillomavirus 1	HPV105	1471	21317	8	3973	5,6,7,8
			HPV12	19	3235	75	280	5,6,7,8
			HPV124	817	75862	0	2769	5,6,8
			HPV14	0	866	1719	0	5,7
			HPV152	4	0	0	0	6
			HPV20	94799	1761	217	111	5,6,7,8

			HPV21	3799	7973	39	1337	5,6,7,8
			HPV24	624	1903	10	568	5,6,7,8
			HPV25	8	0	0	0	6
			HPV36	0	91	1942	196	5,7,8
			HPV47	0	0	12	0	7
			HPV5	7041	2367	38	1026	5,6,7,8
			HPV8	39	0	28	14	6,7,8
			HPV93	46	0	0	0	6
			HPV98	2	6	16	3158	5,6,7,8
		Betapapillomavirus 2	HPV104	4	14	0	0	5,6
			HPV107	259	359	0	21747	5,6,8
			HPV111	8	0	0	12	6,8
			HPV120	302	0	0	20238	6,8
			HPV122	0	0	0	13729	8
			HPV145	0	10	0	0	5
			HPV159	2	7757	0	13692	5,6,8
			HPV17	0	130	0	0	5
			HPV174	17	1146	0	0	5,6
			HPV22	96	15162	0	0	5,6
			HPV23	8	0	0	0	6
			HPV37	458	0	0	26	6,8
			HPV38	18820	22199	1963	0	5,6,7
			HPV80	18	0	0	1357	6,8
			HPV9	2	0	0	0	6
		Betapapillomavirus 3	HPV115	0	0	372	0	7
			HPV49	0	0	1321	0	7
			HPV75	14	114	0	0	5,6
			HPV76	0	0	0	350787	8
		Betapapillomavirus 5	HPV96	502	0	45	0	6,7
	Chipapillomavirus	Chipapillomavirus 2	CPV16	0	0	4	0	7
		Chipapillomavirus 3	CPV10	0	269	0	0	5
	Deltapapillomavirus	Deltapapillomavirus	RalPV1	0	12	0	0	5
		Deltapapillomavirus 2	OvPV1	158	9774	642	3858	5,6,7,8

		Deltapapillomavirus 6	CdPV2	0	0	233	121	7,8
	Dyochipapillomavirus	Dyochipapillomavirus 1	EaPV1	0	884	0	0	5
	Dyoepsilonpapillomavirus	Dyoepsilonpapillomavirus 1	FlPV1	10708	3619	0	1145	5,6,8
	Dyoiotapapillomavirus	Dyoiotapapillomavirus 1	EcPV2	0	0	0	64	8
		Dyoiotapapillomavirus 2	EcPV4	14	0	0	8	6,8
			EcPV5	125	0	0	0	6
	Dyokappapapillomavirus	Dyokappapapillomavirus	BPV16	0	0	27	0	7
			BPV22	0	0	0	22	8
		Dyokappapapillomavirus 1	OaPV3	0	0	2	0	7
	Dyonupapillomavirus	Dyonupapillomavirus 1	ZcPV1	0	0	2	0	7
	Dyoomikronpapillomavirus	Dyoomikronpapillomavirus 1	SscPV3	83	0	0	0	6
	Dyopipapillomavirus	Dyopipapillomavirus 1	PphPV4	0	0	0	1910	8
	Dyorhopapillomavirus	Dyorhopapillomavirus 1	EcPV7	0	0	0	1143	8
	Dyothetapapillomavirus	Dyothetapapillomavirus 1	FcaPV2	0	0	219	0	7
	Dyoxipapillomavirus	Dyoxipapillomavirus 1	BPV7	2	0	0	0	6
	Dyozetapapillomavirus	Dyozetapapillomavirus 1	CcPV1	0	0	10615	0	7
	Epsilonpapillomavirus	Epsilonpapillomavirus 1	BPV5	1954	2961	0	0	5,6
			BPV8	0	0	0	8391	8
	Etapapillomavirus	Etapapillomavirus 1	FcPV1	1195	0	12	16	6,7,8
	Gammapapillomavirus	Gammapapillomavirus	HPV-mCH2	0	0	134	0	7
			HPV-mEV03c05	0	4	0	0	5
			HPV-mEV03c104	25432	0	1072	0	6,7
			HPV-mEV03c212	0	0	0	1488	8
			HPV-mEV03c40	0	48	0	0	5
			HPV-mEV03c434	0	260	0	0	5
			HPV-mEV03c45	0	0	0	4	8
			HPV-mEV07c390	0	0	4	0	7
			HPV-mHIVGc70	0	0	41	0	7
			HPV-mKC5	44	0	724	15	6,7,8
			HPV-mKN3	0	9	0	0	5
			HPV-mMTS2	28	0	0	0	6

			HPV-mTVMBSHc13	0	0	0	2	8
			HPV-mTVMBSWc141	48	0	1511	0	6,7
			HPV-mga2c01	352	0	0	0	6
			HPV-mw07c34d	0	0	2	0	7
			HPV-mw11C39	0	0	45	183898	7,8
			HPV-mw20c04	22	0	0	0	6
			HPV-mw27c39c	0	0	4	1762	7,8
			HPV-mw27c52c	0	2	0	0	5
			HPV-mw34c04a	0	0	0	78713	8
			HPV-mw34c11a	0	0	41	0	7
		Gammapapillomavirus 1	HPV173	0	0	52	0	7
			HPV4	2	42	0	0	5,6
		Gammapapillomavirus 10	HPV121	0	0	29	0	7
			HPV130	0	0	13	0	7
			HPV142	10	0	8295	54504	6,7,8
			HPV180	99	0	0	0	6
		Gammapapillomavirus 11	HPV126	0	2	4	0	5,7
			HPV136	2	0	6	0	6,7
			HPV141	0	0	0	10	8
		Gammapapillomavirus 12	HPV127	0	0	0	66	8
			HPV148	25	0	2	0	6,7
			HPV165	0	0	68	0	7
			HPV199	40	0	0	0	6
		Gammapapillomavirus 13	HPV128	0	0	2	0	7
		Gammapapillomavirus 15	HPV135	0	149656	1383	0	5,7
		Gammapapillomavirus 17	HPV144	0	6	0	0	5
		Gammapapillomavirus 19	HPV162	2	0	2	0	6,7
		Gammapapillomavirus 2	HPV200	0	40	0	0	5
			HPV48	6	0	0	0	6
		Gammapapillomavirus 22	HPVMTS3	2	0	0	0	6
			HPVMTS4	20	0	0	0	6
		Gammapapillomavirus 23	HPV175	0	0	6	0	7
		Gammapapillomavirus 27	HPV201	459	0	0	24	6,8
		Gammapapillomavirus 3	HPV50	0	8	0	0	5

		Gammapapillomavirus 7	HPV109	0	0	14	0	7
			HPV123	0	37	0	0	5
			HPV134	15	4	0	0	5,6
			HPV138	6	28	0	0	5,6
		Gammapapillomavirus 8	HPV119	54	558	0	12650	5,6,8
			HPV147	3402	0	0	0	6
		Gammapapillomavirus 9	HPV129	0	1116	0	0	5
	Kappapapillomavirus	Kappapapillomavirus 2	SfPV1	0	0	0	34024	8
	Lambdapapillomavirus	Lambdapapillomavirus	AmPV4	0	0	2	0	7
	Mupapillomavirus	Mupapillomavirus	HPV-md01c06	0	0	0	18	8
		Mupapillomavirus 1	HPV1	0	0	0	34	8
	Omikronpapillomavirus	Omikronpapillomavirus 1	PphPV1	2	0	0	0	6
	Pipapillomavirus	Pipapillomavirus 1	MaPV1	0	0	0	3799	8
	Psipapillomavirus	Psipapillomavirus	EhPV2	29	0	0	6	6,8
		Psipapillomavirus 1	RaPV1	0	0	0	2	8
	Rhopapillomavirus	Rhopapillomavirus 2	TmPV3	129	0	0	0	6
	Sigmapapillomavirus	Sigmapapillomavirus 1	EdPV1	798	2785	0	7044	5,6,8
	Taupapillomavirus	Taupapillomavirus	MpPV1	0	0	0	242	8
		Taupapillomavirus 1	CPV2	11	0	0	0	6
	Treisdeltapapillomavirus	Treisdeltapapillomavirus 1	RfPV1	0	0	0	2713	8
	Treisepsilonpapillomavirus	Treisepsilonpapillomavirus	PaPV2	363	0	0	419	6,8
		Treisepsilonpapillomavirus 1	PaPV1	0	441	7007	27525	5,7,8
	Treiszetapapillomavirus	Treiszetapapillomavirus	FgPV1	1245	0	1672	48	6,7,8
	Unclassified	Unclassified	EcPV8	10091	0	0	0	6
			MscPV1	0	2	0	0	5
			PpuPV1	0	0	0	2	8
			SaPV1	45855	63157	33903	842220	5,6,7,8
	Upsilonpapillomavirus	Upsilonpapillomavirus 1	TtPV1	0	2	0	0	5
			TtPV3	0	2	10	0	5,7
			TtPV4	1224	0	0	386	6,8
		Upsilonpapillomavirus 2	TtPV2	0	544	0	10	5,8
		Upsilonpapillomavirus 3	PphPV2	0	0	0	25	8
	Xipapillomavirus	Xipapillomavirus	BPV20	91700	124841	00	00	5,6,5
			CePV2					

	Zetapapillomavirus	Zetapapillomavirus 1	EcPV1	0	288	0	0	5

*beta-3-1, beta-3-2, FAP, FAPM1, and FAPM2

Table S3 A: Putative new PV sequences obtained from the NGS analysis.

HPV name	%dissimilarity	Abundance	Number of reads	GInum	Pool	HPV_closest_MegaBlast
2HPVput	11.47	0.0008	20	gi|440573434|gb|KC175574.1|	Pool1_S1_L001	Human papillomavirus isolate FA167 L1 gene, partial cds
3HPVput	10.35	0.0002	6	gi|270048212|gb|FJ947080.1|	Pool1_S1_L001	Human papillomavirus type 115 isolate GC02, complete genome
1HPVput	11.29	0.0021	54	gi|373158195|gb|JN231328.1|	Pool1_S1_L001	Uncultured Papillomavirus contig01 putative L2 and putative L1 genes, complete cds
4HPVput	10.91	0.0001	2	gi|2911551|emb|Y15174.1|	Pool1_S1_L001	Human papillomavirus type 76 E6, E7, E1, E2, E4, L2, and L1 genes
11HPVput	10.64	0.0002	2	gi|1194995709|gb|KY848451.1|	Pool2_S2_L001	Human papillomavirus type 111 isolate HPV/EGY/2015/BI-8 L1 major capsid protein gene, partial cds
14HPVput	11.11	0.0012	12	gi|89574363|gb|DQ418468.1|	Pool2_S2_L001	Human papillomavirus isolate FA152 major capsid protein L1 (L1) gene, partial cds
18HPVput	10.58	0.0002	2	gi|1150189775|gb|KY242581.1|	Pool2_S2_L001	Human papillomavirus isolate EP04 major capsid protein L1 gene, partial cds
19HPVput	10.53	0.0002	2	gi|353441654|gb|JF906528.1|	Pool2_S2_L001	Human papillomavirus isolate SE4 major capsid protein (L1) gene, partial cds
16HPVput	13.51	0.0002	2	gi|1185315494|gb|KY063007.1|	Pool2_S2_L001	Human papillomavirus isolate CT09 major capsid protein (L1) gene, partial cds
12HPVput	11.54	0.0002	2	gi|19387149|gb|AF479251.1|	Pool2_S2_L001	Human papillomavirus isolate FA87 major capsid protein (L1) gene, partial cds
8HPVput	16.31	0.0114	114	gi|28864490|gb|AY204684.1|	Pool2_S2_L001	Human papillomavirus isolate FA101 major capsid protein (L1) gene, partial cds
20HPVput	13.04	0.0002	2	gi|353441652|gb|JF906527.1|	Pool2_S2_L001	Human papillomavirus isolate SE3 major capsid protein (L1) gene, partial cds
5HPVput	15	0.2138	2135	gi|1214938671|gb|MF356498.1|	Pool2_S2_L001	Human papillomavirus isolate ICB1, complete genome
17HPVput	10.17	0.0002	2	gi|1166134776|gb|KY652675.1|	Pool2_S2_L001	Human papillomavirus type 23 strain HPV-23/Lancaster/2015, complete genome
9HPVput	14.1	0.0073	73	gi|11067088|gb|AY009884.1|	Pool2_S2_L001	Human papillomavirus isolate FA49 major capsid protein L1 gene, partial cds
7HPVput	16.13	0.0034	34	gi|39777334|gb|AY468428.1|	Pool2_S2_L001	Human papillomavirus isolate FA131 major capsid protein (L1) gene, partial cds
6HPVput	10.14	0.0827	826	gi|18042187|gb|AF455144.1|	Pool2_S2_L001	Human papillomavirus isolate FA81 major capsid protein (L1) gene, partial cds

10HPVput	10.64	0.0003	3	gi|396918|emb|X74467.1|	Pool2_S2_L001	Human papillomavirus type 14D genomic DNA
13HPVput	12.73	0.0004	4	gi|388771295|gb|JQ963500.1|	Pool2_S2_L001	Human papillomavirus type 120 isolate SIBX3-23 major capture protein L1 (L1) gene, complete cds
15HPVput	12.5	0.0002	2	gi|396940|emb|X74470.1|	Pool2_S2_L001	Human papillomavirus type 19 genomic DNA
21HPVput	10.39	0.0012	11	gi|327195194|gb|JF304769.1|	Pool3_S3_L001	Human papillomavirus type 36 isolate Muc17.2, complete genome
30HPVput	10.87	0.0002	2	gi|1194995707|gb|KY848450.1|	Pool3_S3_L001	Human papillomavirus type 120 isolate HPV/EGY/2015/BI-7 L1 major capsid protein gene, partial cds
29HPVput	10.59	0.0004	4	gi|512390749|gb|KC752132.1|	Pool3_S3_L001	Human papillomavirus isolate KC142 L1 gene, partial cds
36HPVput	10.87	0.0002	2	gi|164564396|gb|EU340869.1|	Pool3_S3_L001	Human papillomavirus isolate FADI1 major capsid protein (L1) gene, partial cds
33HPVput	10.56	0.0002	2	gi|1020186|gb|U31781.1|HPU31781	Pool3_S3_L001	Human papillomavirus type 23, complete genome
28HPVput	14.63	0.0011	10	gi|564732516|gb|KF006398.1|	Pool3_S3_L001	Human papillomavirus type 171, complete genome
38HPVput	17.44	0.0002	2	gi|28864490|gb|AY204684.1|	Pool3_S3_L001	Human papillomavirus isolate FA101 major capsid protein (L1) gene, partial cds
40HPVput	11.25	0.0002	2	gi|870702434|gb|KP692116.1|	Pool3_S3_L001	Human papillomavirus type 202 isolate HPV202, complete genome
26HPVput	13.89	0.0014	13	gi|1144332013|gb|KX781286.1|	Pool3_S3_L001	Human papillomavirus isolate DyskB, complete genome
31HPVput	12.33	0.0035	32	gi|1132312754|gb|KY349817.1|	Pool3_S3_L001	Human papillomavirus isolate MTS1, complete genome
22HPVput	15.15	0.0269	243	gi|512390557|gb|KC752036.1|	Pool3_S3_L001	Human papillomavirus isolate KC34 L1 gene, partial cds
37HPVput	19.23	0.0002	2	gi|15787619|gb|AY049757.1|	Pool3_S3_L001	Human papillomavirus isolate FA66 major capsid protein (L1) gene, partial cds
35HPVput	10.71	0.0002	2	gi|396924|emb|X74468.1|	Pool3_S3_L001	Human papillomavirus type 15 genomic DNA
39HPVput	10.87	0.0007	6	gi|238623442|emb|FM955839.1|	Pool3_S3_L001	Human papillomavirus type 100, complete genome
23HPVput	19.48	0.0022	20	gi|293596086|gb|HM011570.1|	Pool3_S3_L001	Gammapapillomavirus HPV127 isolate R3a, complete genome
25HPVput	12.64	0.0039	35	gi|39777334|gb|AY468428.1|	Pool3_S3_L001	Human papillomavirus isolate FA131 major capsid protein (L1) gene, partial cds
27HPVput	13.51	0.0013	12	gi|39777351|gb|AY468436.1|	Pool3_S3_L001	Human papillomavirus isolate FA20.4 major capsid protein (L1) gene, partial cds

41HPVput	11.59	0.0002	2	gi|89574381|gb|DQ418477.1|	Pool3_S3_L001	Human papillomavirus isolate FA162 major capsid protein L1 (L1) gene, partial cds
24HPVput	13.33	0.0008	7	gi|530669309|emb|HG421739.1|	Pool3_S3_L001	Human papillomavirus type 179 complete genome, isolate SIBX16
34HPVput	21.43	0.0004	4	gi|28864508|gb|AY204693.1|	Pool3_S3_L001	Human papillomavirus isolate FA109 major capsid protein (L1) gene, partial cds
42HPVput	11.43	0.0002	2	gi|312451781|gb|GU117629.1|	Pool3_S3_L001	Human papillomavirus type 149, complete genome
32HPVput	14.75	0.0004	4	gi|356483524|gb|JN171845.1|	Pool3_S3_L001	Human Papillomavirus type 153, complete genome
57HPVput	10.71	0.0008	7	gi|327195194|gb|JF304769.1|	Pool4_S4_L001	Human papillomavirus type 36 isolate Muc17.2, complete genome
76HPVput	13.24	0.0009	8	gi|380865534|gb|JQ250749.1|	Pool4_S4_L001	Human papillomavirus isolate SE53 L1 gene, partial cds
71HPVput	10.53	0.0002	2	gi|1020234|gb|U31787.1|HPU31787	Pool4_S4_L001	Human papillomavirus type 38, complete genome
69HPVput	10.59	0.002	18	gi|6694870|gb|AF217656.1|AF217656	Pool4_S4_L001	Human papillomavirus isolate FA14 major capsid protein L1 gene, partial cds
80HPVput	17.44	0.0002	2	gi|6694910|gb|AF217676.1|AF217676	Pool4_S4_L001	Human papillomavirus isolate FA31 major capsid protein L1 gene, partial cds
55HPVput	10.42	0.0003	3	gi|19423659|gb|AF489710.1|	Pool4_S4_L001	Human papillomavirus isolate FAIMVS6.3 major capsid protein (L1) gene, partial cds
49HPVput	12.25	0.0043	38	gi|15020297|gb|AY040281.1|	Pool4_S4_L001	Human papillomavirus isolate FA61 major capsid protein (L1) gene, partial cds
45HPVput	14.07	0.0414	368	gi|89574363|gb|DQ418468.1|	Pool4_S4_L001	Human papillomavirus isolate FA152 major capsid protein L1 (L1) gene, partial cds
81HPVput	10.22	0.0002	2	gi|389885562|gb|JN211195.1|	Pool4_S4_L001	Human papillomavirus type 17 isolate S410, complete genome
48HPVput	10.2	0.0018	16	gi|1150189775|gb|KY242581.1|	Pool4_S4_L001	Human papillomavirus isolate EP04 major capsid protein L1 gene, partial cds
82HPVput	15.83	0.0002	2	gi|255683772|gb|FJ492744.1|	Pool4_S4_L001	Canine papillomavirus 6 isolate Zurich, complete genome
65HPVput	14.95	0.0029	26	gi|1020258|gb|U31790.1|HPU31790	Pool4_S4_L001	Human papillomavirus type 50, complete genome
74HPVput	11.29	0.0002	2	gi|512390537|gb|KC752026.1|	Pool4_S4_L001	Human papillomavirus isolate KC24 L1 gene, partial cds
77HPVput	10.47	0.0004	4	gi|238623458|emb|FM955841.1|	Pool4_S4_L001	Human papillomavirus type 105, complete genome
68HPVput	13.71	0.0009	8	gi|6694896|gb|AF217669.1|AF217669	Pool4_S4_L001	Human papillomavirus isolate FA24.2 major capsid protein L1 gene, partial cds

43H PVput	14.17	0.0689	613	gi|189003642|gb|EU541441.1|	Pool4_S4_L001	Human papillomavirus type 109, complete genome
83H PVput	13.67	0.0002	2	gi|421991531|gb|JQ612578.1|	Pool4_S4_L001	Human papillomavirus isolate GC21 major capsid protein L1 gene, partial cds
64H PVput	10.07	0.0034	30	gi|50236486|gb|AY382779.2|	Pool4_S4_L001	Human papillomavirus type 96, complete genome
53H PVput	11.96	0.0009	8	gi|343411569|gb|HM999994.1|	Pool4_S4_L001	Human papillomavirus type 142 isolate GH1302, complete genome
46H PVput	18.08	0.0794	706	gi|28864490|gb|AY204684.1|	Pool4_S4_L001	Human papillomavirus isolate FA101 major capsid protein (L1) gene, partial cds
78H PVput	11.7	0.0022	20	gi|49425436|gb|AY468429.2|	Pool4_S4_L001	Human papillomavirus isolate FA132 major capsid protein (L1) gene, partial cds
54H PVput	11.67	0.0015	13	gi|270048230|gb|FJ969899.1|	Pool4_S4_L001	Human papillomavirus isolate GC07_1 major capsid protein L1 gene, partial cds
61H PVput	10.17	0.0007	6	gi|2894523|emb|AJ223858.1|	Pool4_S4_L001	human papillomavirus type 24, L1 capsid gene strain HPV24
67H PVput	18.9	0.0002	2	gi|1144331551|gb|KX781282.1|	Pool4_S4_L001	Human papillomavirus isolate Dysk3, complete genome
73H PVput	12.61	0.0031	28	gi|6694920|gb|AF217681.1|AF217681	Pool4_S4_L001	Human papillomavirus isolate FA36 major capsid protein L1 gene, partial cds
84H PVput	10.17	0.0002	2	gi|270048242|gb|FJ969905.1|	Pool4_S4_L001	Human papillomavirus isolate GC10 major capsid protein L1 gene, partial cds
59H PVput	17.95	0.0038	34	gi|296495828|gb|GQ845441.1|	Pool4_S4_L001	Human papillomavirus type 119, complete genome
75H PVput	13.56	0.0011	10	gi|396924|emb|X74468.1|	Pool4_S4_L001	Human papillomavirus type 15 genomic DNA
72H PVput	12.79	0.0022	20	gi|929996745|gb|KT372348.1|	Pool4_S4_L001	Human papillomavirus type 199 isolate A484-SLO, complete genome
85H PVput	10.83	0.0002	2	gi|270048212|gb|FJ947080.1|	Pool4_S4_L001	Human papillomavirus type 115 isolate GC02, complete genome
70H PVput	10.17	0.0004	4	gi|270048238|gb|FJ969903.1|	Pool4_S4_L001	Human papillomavirus isolate GC09 major capsid protein L1 gene, partial cds
79H PVput	12.99	0.0011	10	gi|1020202|gb|U31783.1|HPU31783	Pool4_S4_L001	Human papillomavirus type 28, complete genome
58H PVput	11.2	0.0046	41	gi|45184601|gb|AY546078.1|	Pool4_S4_L001	Human papillomavirus isolate FA140.2 major capsid protein L1 gene, partial cds
44H PVput	11.24	0.0572	509	gi|270048254|gb|FJ969911.1|	Pool4_S4_L001	Human papillomavirus isolate GC14 major capsid protein L1 gene, partial cds
66H PVput	10.58	0.0002	2	gi|22074072|gb|AY044276.1|	Pool4_S4_L001	Human papillomavirus isolate P36-14 major capsid protein L1 (L1) gene, partial cds

63HPVput	12.39	0.004	36	gi|293596086|gb|HM011570.1|	Pool4_S4_L001	Gammapapillomavirus HPV127 isolate R3a, complete genome
50HPVput	11.59	0.0239	213	gi|537801747|gb|KF482069.1|	Pool4_S4_L001	Human papillomavirus isolate HPV-L55, complete genome
60HPVput	23.27	0.0004	4	gi|1185315490|gb|KY063005.1|	Pool4_S4_L001	Human papillomavirus isolate CT07 major capsid protein (L1) gene, partial cds
52HPVput	13.04	0.0066	59	gi|89574379|gb|DQ418476.1|	Pool4_S4_L001	Human papillomavirus isolate FA160 major capsid protein L1 (L1) gene, partial cds
56HPVput	11.94	0.0006	5	gi|19423667|gb|AF489714.1|	Pool4_S4_L001	Human papillomavirus isolate FAIMVS9 major capsid protein (L1) gene, partial cds
47HPVput	10.53	0.0016	14	gi|18042187|gb|AF455144.1|	Pool4_S4_L001	Human papillomavirus isolate FA81 major capsid protein (L1) gene, partial cds
62HPVput	12.16	0.0013	12	gi|528225988|gb|KC878009.1|	Pool4_S4_L001	Human papillomavirus isolate SE125 L1 (L1) gene, partial cds
51HPVput	10.64	0.0016	14	gi|1194995705|gb|KY848449.1|	Pool4_S4_L001	Human papillomavirus type 21 isolate HPV/EGY/2015/BI-1 L1 major capsid protein gene, partial cds
88HPVput	11.02	0.0005	6	gi|1185315498|gb|KY063009.1|	Pool5_S5_L001	Human papillomavirus isolate CT11 major capsid protein (L1) gene, partial cds
91HPVput	10.84	0.0003	4	gi|2894523|emb|AJ223858.1|	Pool5_S5_L001	human papillomavirus type 24, L1 capsid gene strain HPV24
87HPVput	10.91	0.0062	73	gi|1132312754|gb|KY349817.1|	Pool5_S5_L001	Human papillomavirus isolate MTS1, complete genome
89HPVput	13.54	0.011	130	gi|395627595|emb|HE963025.1|	Pool5_S5_L001	Human papillomavirus type 159 complete genome, isolate SIBX8
90HPVput	19.05	0.0002	2	gi|353441684|gb|JF906543.1|	Pool5_S5_L001	Human papillomavirus isolate SE23 major capsid protein (L1) gene, partial cds
92HPVput	10.26	0.0007	8	gi|396918|emb|X74467.1|	Pool5_S5_L001	Human papillomavirus type 14D genomic DNA
86HPVput	18.39	0.5125	6080	gi|353441668|gb|JF906535.1|	Pool5_S5_L001	Human papillomavirus isolate SE13 major capsid protein (L1) gene, partial cds
95HPVput	11.91	0.0001	2	gi|40686789|gb|AY502598.1|	Pool6_S6_L001	Human papillomavirus isolate FA141 major capsid protein L1 gene, partial cds
93HPVput	11.24	0.0004	5	gi|74484000|gb|DQ090005.2|	Pool6_S6_L001	Human papillomavirus type 38b subtype FA125, complete genome
94HPVput	18.39	0.001	14	gi|353441668|gb|JF906535.1|	Pool6_S6_L001	Human papillomavirus isolate SE13 major capsid protein (L1) gene, partial cds
96HPVput	12.16	0.0003	4	gi|89574371|gb|DQ418472.1|	Pool7_S7_L001	Human papillomavirus isolate FA156 major capsid protein L1 (L1) gene, partial cds
97HPVput	11.24	0.0001	2	gi|270048254|gb|FJ969911.1|	Pool7_S7_L001	Human papillomavirus isolate GC14 major capsid protein L1 gene, partial cds

104HPVput	11.25	0.0001	2	gi|1200175664|gb|KY969593.1|	Pool8_S8_L001	Human papillomavirus type 20 strain HPV-20/Lancaster/2015, complete genome
98HPVput	12	0.113	2713	gi|1150189779|gb|KY242583.1|	Pool8_S8_L001	Human papillomavirus type 209 isolate FA108, complete genome
102HPVput	11.67	0.0002	6	gi|390517205|emb|HE820175.1|	Pool8_S8_L001	Human papillomavirus type 42 L1 gene for L1 protein, isolate A207
101HPVput	10.67	0.0003	8	gi|353441702|gb|JF906553.1|	Pool8_S8_L001	Human papillomavirus isolate SE36 major capsid protein (L1) gene, partial cds
105HPVput	11.43	0.0001	2	gi|49425436|gb|AY468429.2|	Pool8_S8_L001	Human papillomavirus isolate FA132 major capsid protein (L1) gene, partial cds
99HPVput	11.04	0.1378	3308	gi|1132312754|gb|KY349817.1|	Pool8_S8_L001	Human papillomavirus isolate MTS1, complete genome
100HPVput	11.54	0.0007	18	gi|395627595|emb|HE963025.1|	Pool8_S8_L001	Human papillomavirus type 159 complete genome, isolate SIBX8
103HPVput	12.25	0.0001	2	gi|388771295|gb|JQ963500.1|	Pool8_S8_L001	Human papillomavirus type 120 isolate SIBX3-23 major capture protein L1 (L1) gene, complete cds

*beta-3-1, beta-3-2, FAP, FAPM1, and FAPM2

Table S3 B: Putative new PV sequences obtained from the NGS analysis.

HPVname	Specimen	Primer	Length	HPV_closest_PaVE_Blast	AlignmentPosition_start:stop(length)	Classification_PaVE	Classification_RAxML
2HPVput	skin	Beta-3-1/Beta-3-2	138/49	HPV115(89.57%)/HPV76(91.84%)	1-115(115)/1-49(49)	Betapapillomavirus 3/Betapapillomavirus 3	Betapapillomavirus 3/Betapapillomavirus 3
3HPVput	skin	Beta-3-1/Beta-3-2	58	HPV115(89.66%)	1-58(58)	Betapapillomavirus 3	Dyophipapillomavirus 1
1HPVput	skin	Beta-3-1/Beta-3-2	66/162	HPV174(85.94%)/HPV49(89.51%)	1-64(64)/1-162(162)	Betapapillomavirus 2/Betapapillomavirus 3	Psipapillomavirus/Unclassified
4HPVput	skin	Beta-3-1/Beta-3-2	63	HPV76(89.09%)	9-63(55)	Betapapillomavirus 3	Betapapillomavirus 3
11HPVput	skin	FAP	47	HPV113(89.36%)	1-47(47)	Betapapillomavirus 2	Betapapillomavirus 2
14HPVput	skin	FAP	72	HPV164(80.28%)	1-71(71)	Gammapapillomavirus 8	Gammapapillomavirus
18HPVput	skin	FAP	104	HPV142(94.23%)	1-104(104)	Gammapapillomavirus 10	Gammapapillomavirus 10
19HPV	skin	FAP	96	HPV128(81.18%)	7-90(84)	Gammapapillomavirus 13	Gammapapillomavirus 13

Vput16HP	skin	FAP	171	HPV65(80.86%)	2-163(162)	Gammapapillomavirus 1	Gammapapillomavirus 1
Vput12HP	skin	FAP	67	HPV170(92.86%)	10-65(56)	Gammapapillomavirus 7	Gammapapillomavirus
PVput20HP	skin	FAP	139/100	HPV123(91.37%)/HPV123(92%)	1-139(139)/1-100(100)	Gammapapillomavirus 7/Gammapapillomavirus 7	Unclassified/Gammapapillomavirus 7
Vput5HPV	skin	FAP	215	HPV_MTS2(77.21%)	1-215(215)	Gammapapillomavirus 7	Unclassified
PVput17HP	skin	FAP	67	HPV148(86.57%)	1-67(67)	Gammapapillomavirus 12	Gammapapillomavirus 12
Vput9HPV	skin	FAP	59	HPV23(89.83%)	1-59(59)	Betapapillomavirus 2	Betapapillomavirus 2
PVput7HPV	skin	FAP	89	HPV138(92.13%)	1-89(89)	Gammapapillomavirus 7	Gammapapillomavirus 7
PVput6HPV	skin	FAP	69/87	HPV99(85.51%)/HPV19(86.05%)	1-69(69)/1-86(86)	Betapapillomavirus 1/Betapapillomavirus 1	Betapapillomavirus 1/Betapapillomavirus 2
PVput10HP	skin	FAP	69	HPV148(80.88%)	2-69(68)	Gammapapillomavirus 12	Taupapillomavirus 3
Vput13HP	skin	FAP	47	HPV25(89.36%)	1-47(47)	Betapapillomavirus 1	Dyorhopapillomavirus 1
Vput15HP	skin	FAP	56/58	HPV120(87.27%)/HPV120(87.93%)	2-56(55)/1-58(58)	Betapapillomavirus 2/Betapapillomavirus 2	Betapapillomavirus/Deltapapillomavirus 5
Vput21HP	skin	FAP	72	HPV38(96.3%)	19-72(54)	Betapapillomavirus 2	Betapapillomavirus 1
Vput30HP	skin	FAPM1	77/140	HPV36(89.61%)/HPV36(85.71%)	1-77(77)/1-140(140)	Betapapillomavirus 1/Betapapillomavirus 1	Treiszetapapillomavirus/Betapapillomavirus
Vput29HP	skin	FAPM1	46	HPV120(89.13%)	1-46(46)	Betapapillomavirus 2	Betapapillomavirus 2
Vput36HP	skin	FAPM1	118	HPV116(80.37%)	10-116(107)	Gammapapillomavirus 9	Unclassified
Vput33HP	skin	FAPM1	53	HPV22(97.87%)	1-47(47)	Betapapillomavirus 2	Betapapillomavirus 2
Vput28HP	skin	FAPM1	190	HPV23(89.44%)	49-190(142)	Betapapillomavirus 2	Unclassified
Vput38HP	skin	FAPM1	136	HPV4(78.42%)	1-136(136)	Gammapapillomavirus 1	Gammapapillomavirus
Vput40HP	skin	FAPM1	86	HPV123(91.86%)	1-86(86)	Gammapapillomavirus 7	Gammapapillomavirus 7
Vput26HP	skin	FAPM1	80	HPV202(88.75%)	1-80(80)	Gammapapillomavirus 11	Gammapapillomavirus 11
	skin	FAPM1	76	HPV37(85.92%)	6-76(71)	Betapapillomavirus 2	Deltapapillomavirus 2

Vput31							
HPVput22	skin	FAPM1	145	HPV_MTS1(86.9%)	1-145(145)	Betapapillomavirus 2	Betapapillomavirus 2
HPVput37	skin	FAPM1	99	HPV37(83.84%)	1-99(99)	Betapapillomavirus 2	Betapapillomavirus 2
HPVput35	skin	FAPM1	197	HPV126(78.97%)	1-192(192)	Gammapapillomavirus 11	Gammapapillomavirus
HPVput39	skin	FAPM1	142	HPV15(89.29%)	3-142(140)	Betapapillomavirus 2	Betapapillomavirus 2
HPVput23	skin	FAPM1	93	HPV100(89.13%)	1-92(92)	Betapapillomavirus 2	Unclassified
HPVput25	skin	FAPM1	111	HPV126(85.59%)	1-111(111)	Gammapapillomavirus 11	Gammapapillomavirus 11
HPVput27	skin	FAPM1	88	HPV37(85.06%)	2-88(87)	Betapapillomavirus 2	Deltapapillomavirus 2
HPVput41	skin	FAPM1	128	HPV15(93.6%)	2-126(125)	Betapapillomavirus 2	Gammapapillomavirus 9
HPVput24	skin	FAPM1	74	HPV17(88.89%)	7-69(63)	Betapapillomavirus 2	Betapapillomavirus 2
HPVput34	skin	FAPM1	88	HPV166(81.48%)	1-81(81)	Gammapapillomavirus 19	Lambdapapillomavirus 3
HPVput42	skin	FAPM1	223	HPV149(82.24%)	11-223(213)	Gammapapillomavirus 7	Unclassified
HPVput32	skin	FAPM1	107	HPV149(87.5%)	1-104(104)	Gammapapillomavirus 7	Unclassified
HPVput57	skin	FAPM1	75	HPV144(81.08%)	2-75(74)	Gammapapillomavirus 17	Deltapapillomavirus 2
HPVput76	skin	CUT	57	HPV24(91.23%)	1-57(57)	Betapapillomavirus 1	Treiszetapapillomavirus
HPVput71	skin	CUT	107	HPV92(82.08%)	1-106(106)	Betapapillomavirus 4	Betapapillomavirus
HPVput69	skin	CUT	92	HPV38(90.91%)	1-85(85)	Betapapillomavirus 2	Unclassified
HPVput80	skin	CUT	85	HPV17(90.91%)	1-66(66)	Betapapillomavirus 2	Betapapillomavirus
HPVput55	skin	CUT	94	HPV148(76.6%)	1-94(94)	Gammapapillomavirus 12	Gammapapillomavirus 12
HPVput49	skin	CUT	48	HPV5(89.58%)	1-48(48)	Betapapillomavirus 1	Betapapillomavirus 1
HP	skin	CUT	98	HPV156(76.84%)	1-95(95)	Gammapapillomavirus 18	Gammapapillomavirus

Vput 45 HP							
Vput 81 HP	skin	CUT	147	HPV60(73.86%)	1-147(147)	Gammapapillomavirus 4	Gammapapillomavirus
Vput 48 HP	skin	CUT	143	HPV17(90.3%)	10-143(134)	Betapapillomavirus 2	Dyoomikronpapillomavirus 1
Vput 82 HP	skin	CUT	109	HPV142(94.5%)	1-109(109)	Gammapapillomavirus 10	Gammapapillomavirus 10
Vput 65 HP	skin	CUT	208	CPV6(75.7%)	1-208(208)	Lambdapapillomavirus 3	Lambdapapillomavirus 3
Vput 74 HP	skin	CUT	127/123	HPV168(85.25%)/HPV50(88.03%)	1-122(122)/7-123(117)	Gammapapillomavirus 8/Gammapapillomavirus 3	Gammapapillomavirus/Gammapapillomavirus 21
Vput 77 HP	skin	CUT	67	HPV98(95.38%)	1-65(65)	Betapapillomavirus 1	Deltapapillomavirus 2
Vput 68 HP	skin	CUT	86	HPV105(89.53%)	1-86(86)	Betapapillomavirus 1	Dyorhopapillomavirus 1
Vput 43 HP	skin	CUT	128	HPV136(83%)	21-120(100)	Gammapapillomavirus 11	Unclassified
Vput 83 HP	skin	CUT	134	HPV123(93.28%)	1-119(119)	Gammapapillomavirus 7	Gammapapillomavirus 7
Vput 64 HP	skin	CUT	117	HPV200(76.92%)	1-117(117)	Gammapapillomavirus 2	Gammapapillomavirus 14
Vput 53 HP	skin	CUT	144	HPV96(89.93%)	1-139(139)	Betapapillomavirus 5	Betapapillomavirus 5
Vput 46 HP	skin	CUT	92	HPV142(88.04%)	1-92(92)	Gammapapillomavirus 10	Gammapapillomavirus 10
Vput 78 HP	skin	CUT	111	HPV123(90.91%)	1-55(55)	Gammapapillomavirus 7	Gammapapillomavirus 7
Vput 54 HP	skin	CUT	112/120	HPV118(81.73%)/HPV36(87.07%)	1-104(104)/1-116(116)	Betapapillomavirus 1/Betapapillomavirus 1	Unclassified/Unclassified
Vput 61 HP	skin	CUT	60	HPV180(84.75%)	2-60(59)	Gammapapillomavirus 10	Chipapillomavirus 3
Vput 67 HP	skin	CUT	64	HPV24(89.83%)	6-64(59)	Betapapillomavirus 1	Betapapillomavirus 3
Vput 73 HP	skin	CUT	136	HPV123(93.02%)	1-129(129)	Gammapapillomavirus 7	Gammapapillomavirus 7
Vput 84 HP	skin	CUT	120	HPV36(85.83%)	1-120(120)	Betapapillomavirus 1	Gammapapillomavirus 11
Vput 59 HP	skin	CUT	59	HPV77(87.88%)	21-53(33)	Alphapapillomavirus 2	Alphapapillomavirus 3
	skin	CUT	198	HPV119(82.91%)	1-198(198)	Gammapapillomavirus 8	Gammapapillomavirus

Vput75HP							
Vput72HP	skin	CUT	95	HPV15(86.44%)	1-59(59)	Betapapillomavirus 2	Gammapapillomavirus
Vput85HP	skin	CUT	88/99	HPV199(86.42%)/HPV123(90.22%)	8-88(81)/1-92(92)	Gammapapillomavirus 12/Gammapapillomavirus 7	Chipapillomavirus 2/Gammapapillomavirus 7
Vput70HP	skin	CUT	120	HPV115(89.17%)	1-120(120)	Betapapillomavirus 3	Betapapillomavirus 3
Vput79HP	skin	CUT	59	HPV165(80.36%)	4-59(56)	Gammapapillomavirus 12	Gammapapillomavirus
Vput58HP	skin	CUT	83	HPV28(85.54%)	1-83(83)	Alphapapillomavirus 2	Alphapapillomavirus 2
Vput44HP	skin	CUT	125	HPV17(88%)	1-125(125)	Betapapillomavirus 2	Betapapillomavirus 2
Vput66HP	skin	CUT	89	HPV36(77.97%)	30-88(59)	Betapapillomavirus 1	Unclassified
Vput63HP	skin	CUT	105	HPV36(87.5%)	2-105(104)	Betapapillomavirus 1	Unclassified
Vput50HP	skin	CUT	134	HPV127(84.21%)	1-133(133)	Gammapapillomavirus 12	Gammapapillomavirus 12
Vput60HP	skin	CUT	106	HPV172(77.91%)	19-104(86)	Gammapapillomavirus 22	Gammapapillomavirus
Vput52HP	skin	CUT	256	HPV168(72.73%)	1-250(250)	Gammapapillomavirus 8	Unclassified
Vput56HP	skin	CUT	105/62/77/101	HPV115(83%)/HPV119(88.71%)/HPV168(83.33%)/HPV113(78.49%)	1-99(99)/1-62(62)/6-71(66)/2-94(93)	Betapapillomavirus 3/Gammapapillomavirus 8/Gammapapillomavirus 8/Betapapillomavirus 2	Lambdapapillomavirus 5/Gammapapillomavirus 8/Deltapapillomavirus 2/Gammapapillomavirus 8
Vput47HP	skin	CUT	67/265	HPV147(86.05%)/HPV168(73.48%)	1-43(43)/2-265(264)	Gammapapillomavirus 8/Gammapapillomavirus 8	Gammapapillomavirus/Unclassified
Vput62HP	skin	CUT	57	HPV157(94.23%)	6-57(52)	Gammapapillomavirus	Dyoepsilonpapillomavirus 1
Vput51HP	skin	CUT	76	HPV76(86.49%)	3-76(74)	Betapapillomavirus 3	Nupapillomavirus 1
Vput88HP	skin	CUT	47	HPV21(89.36%)	1-47(47)	Betapapillomavirus 1	Betapapillomavirus 1
Vput91HP	oral	FAPM2	129	HPV154(79.59%)	7-104(98)	Gammapapillomavirus 11	Gammapapillomavirus
Vput87HP	oral	FAPM2	165	HPV24(89.16%)	1-165(165)	Betapapillomavirus 1	Betapapillomavirus 1
Vput89HP	oral	FAPM2	60	HPV22(93.22%)	2-60(59)	Betapapillomavirus 2	Betapapillomavirus 2
	oral	FAPM2	98	HPV22(89.13%)	1-92(92)	Betapapillomavirus 2	Betapapillomavirus 2

Vput 90 HP							
Vput 92 HP	oral	FAPM 2	94	HPV24(91.3%)	3-94(92)	Betapapillomavirus 1	Gammapapillomavirus 8
Vput 86 HP	oral	FAPM 2	126	HPV14(88.1%)	1-126(126)	Betapapillomavirus 1	Betapapillomavirus 1
Vput 95 HP	oral	FAPM 2	118	HPV22(92.37%)	1-118(118)	Betapapillomavirus 2	Betapapillomavirus 2
Vput 93 HP	oral	FAPM 1	84	HPV21(92.86%)	1-84(84)	Betapapillomavirus 1	Betapapillomavirus 1
Vput 94 HP	oral	FAPM 1	92	HPV_MTS1(92.21%)	1-77(77)	Betapapillomavirus 2	Sigmapapillomavirus 1
Vput 96 HP	oral	FAPM 1	144	HPV22(92.36%)	1-144(144)	Betapapillomavirus 2	Dyoiotapapillomavirus 2
Vput 97 HP	oral	CUT	74	HPV36(86.49%)	1-74(74)	Betapapillomavirus 1	Chipapillomavirus 2
Vput 104 HP	oral	CUT	89	HPV36(77.97%)	30-88(59)	Betapapillomavirus 1	Unclassified
Vput 98 HP	oral	MIX*	80	HPV20(88.75%)	1-80(80)	Betapapillomavirus 1	Treisepsilonpapillomavirus
Vput 102 HP	oral	MIX	97	HPV_MTS1(87.63%)	1-97(97)	Betapapillomavirus 2	Treisdeltapapillomavirus 1
Vput 101 HP	oral	MIX	60	HPV49(86.67%)	1-60(60)	Betapapillomavirus 3	Treisepsilonpapillomavirus 1
Vput 105 HP	oral	MIX	85	HPV171(88.89%)	5-76(72)	Gammapapillomavirus 11	Treiszetapapillomavirus
Vput 99 HP	oral	MIX	67	HPV20(100%)	28-67(40)	Betapapillomavirus 1	Betapapillomavirus
Vput 100 HP	oral	MIX	163	HPV_MTS1(88.96%)	1-163(163)	Betapapillomavirus 2	Unclassified
Vput 103 HP	oral	MIX	52	HPV159(88.46%)	1-52(52)	Betapapillomavirus 2	Mupapillomavirus
Vput	oral	MIX	49	HPV120(87.76%)	1-49(49)	Betapapillomavirus 2	Betapapillomavirus