From DEPARTMENT OF LABORATORY MEDICINE
Karolinska Institutet, Stockholm, Sweden

INFECTIONS IN SKIN CANCER

Laila Sara Arroyo Mühr

Stockholm 2016

Cover photography: HPV 197 L1 protein. Predicted by The Phyre2 web portal for protein modeling, prediction and analysis. Kelley LA et al. Nature Protocols 10, 845-858 (2015).

Published by Karolinska Institutet

Printed by Eprint AB 2016

ISBN 978-91-7676-215-8

Department of Laboratory Medicine

Infections in Skin Cancer

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Föreläsningssal Solen 4U, Alfred Nobels Allé 8, Karolinska Institutet, Huddinge.

Fredagen den 08 april 2016, kl 13.00

av
Laila Sara Arroyo Mühr
MSc Pharmacy

Principal Supervisor:
Professor Joakim Dillner
Karolinska Institutet
Department of Laboratory Medicine
Division of Pathology

Co-supervisor(s):
PhD Emilie Hultin
Karolinska Institutet
Department of Laboratory Medicine
Division of Pathology

Associate Professor Ola Forslund
Lund University
Department of Laboratory Medicine
Division of Medical Microbiology

Professor Göran Andersson
Karolinska Institutet
Department of Laboratory Medicine
Division of Pathology

Opponent:
PhD Max Käller
Royal Institute of Technology
Division of Gene Technology

Examination Board:
Professor Ingemar Ernberg
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Examination Board:
Professor Lars Engstrand
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Examination Board:
Professor Emeritus Jonas Blomberg
Uppsala University
Department of Medical Science

Stockholm 2016

“We only see what we know”

(J.W. von Goethe)

ABSTRACT

The increasing prevalence of skin cancer results in that it will soon equal that of all other cancers combined. Sun exposure is a well-known risk factor for its development, but despite the growing public awareness of the harmful consequences of ultraviolet radiation, the cancer incidence continues to increase, implying that other factors might also have a role in promoting this disease.

Data from immunosuppressed patients reveals a 100-fold increased incidence of non-melanoma skin carcinoma (NMSC), but an infectious etiology has not been established. However, certain human papillomaviruses (HPVs) have previously been detected in this type of cancer.

We applied high throughput sequencing to different skin lesions in order to assess which organisms were present. Most viral reads (>95%) belonged to human papillomavirus.

Traditionally, viral detection was performed using PCR methods. We used degenerate "general" HPV primers and multiplexed novel "specific" HPV primers in order to amplify a broad number of HPVs by PCR. This method showed a very high sensitivity, but the HPV types with low similarity to the primer sequences might have escaped amplification. Therefore, we performed an unbiased approach based on non-PCR whole genome amplification, independent of sequence information, in order to detect those "escaping" HPV types, as well as to determine if other viruses were present in the samples.

Overall, we identified almost 100 putative novel HPV types in total, and characterized 4 novel HPV types (HPV 197, 200, 201 and 202). Most of the HPV types were detected in very few patients each, and at a very low viral load (below 0.5 copies/cell), except for HPV 197, which was the most commonly found virus in skin tumors (37.4% of skin lesions). Despite the higher sensitivity of PCR methods, the unbiased approach detected HPV in 37/40 condyloma acuminata that had been reported as "HPV-negative" with specific PCR techniques. Certain HPV types, including HPV 197, were not detected by PCR and only by non-PCR based methods. Therefore, more unbiased PCR-independent methods are needed to describe which organisms are most commonly present in skin lesions.

The work in this thesis has expanded our knowledge of the wide genomic diversity of HPV on the skin, and finds that PCR-independent methods are needed to describe which organisms are most commonly present in skin lesions. Further studies are needed to assess any possible role of viral infections in skin cancer, elucidation of mechanistic effects and determine the direction of causality of any associations.

SAMMANFATTNING

Den ökande förekomsten av hudcancer resulterar snart i lika många fall som alla andra cancertyper tillsammans. Solexponering är en känd riskfaktor för utveckling av hudcancer, men trots allmän kännedom kring de skadliga konsekvenserna av ultraviolett strålning, så ökar frekvensen av sjukdomen, vilket tyder på att det även kan finnas andra bidragande faktorer.

Sjukdomsstatistik från patienter med nedsatt immunförsvar visar en 100-faldig ökad frekvens av icke-melanom hudcancer, men någon bakomliggande infektion har ännu inte kunnat fastställas. Dock har vissa typer av humant papillomvirus (HPV) hittats i denna typ av cancer.

Vi sekvenserade allt DNA i olika hudförändringar för att undersöka vilka mikroorganismer de innehöll. De flesta virussekvenserna (>95%) kom från HPV.

Traditionellt har virus detekterats med olika PCR-metoder. Vi använde oss av degenererade ”generella” HPV-primers och multiplexade nya ”specifika” HPV-primers för att möjliggöra PCR-amplifiering av många olika HPV-typer. Denna metod visade på en mycket hög känslighet, men HPV-typer med låg likhet till primersekvenserna kan ha undgått amplifiering. För ett mer objektivt tillvägagångssätt amplifierade vi allt DNA utan PCR och oberoende av någon sekvensinformation för att kunna detektera eventuella HPV-typer som kan ha undgått PCR-amplifieringen likväl som andra virus i proverna.

Totalt identifierade vi nära 100 möjliga nya HPV-typer, samt karaktäriserade 4 nya HPV-typer (HPV 197, 200, 201 och 202). De flesta HPV-typerna detekterades bara i några få patienter var, med mycket låga virustal (mindre än 0,5 kopior/cell), förutom HPV 197, vilket var det vanligast förekommande viruset bland hudtumörer (37,4% av hudförändringarna). Trots den högre känsligheten hos PCR-baserade metoder, detekterade den mer objektiva PCR-oberoende metoden HPV i 37/40 condylomata acuminata som alla tidigare rapporterats som HPV-negativa med specifika PCR-metoder. Vissa HPV-typer, inklusive HPV 197, detekterades inte med PCR, utan enbart med metoder utan PCR. Därför behövs fler objektiva, PCR-oberoende, metoder för att beskriva vilka mikroorganismer som är vanligast förekommande i hudförändringar.

Arbetet i denna avhandling har utökat vår kunskap om den stora genetiska mångfalden av HPV i hud, samt konstaterar att PCR-oberoende metoder är nödvändiga för att beskriva vilka mikroorganismer som är vanligast förekommande i hudförändringar. Vidare studier är nödvändiga för att fastställa möjliga samband mellan virusinfektioner och hudcancer, klargöra mekanistiska effekter, samt avgöra orsaksriktning mellan funna samband.

RESUMEN

El cáncer de piel es el más frecuente de los cánceres en el ser humano. A pesar de que la exposición a la luz ultravioleta es un factor de riesgo bien conocido y de la creciente concienciación popular sobre sus efectos perjudiciales, la incidencia de este cáncer continúa aumentando. Se estima que no tardará mucho en sobrepasar en número a la suma del resto de cánceres. Esto sugiere que pueden existir otros factores que contribuyen al desarrollo de esta enfermedad.

Las personas inmunodeprimidas presentan una mayor incidencia en la mayoría de los cánceres, sobre todo en los causados por virus oncogénicos (consecuencia de la reducción general de su respuesta inmune). El cáncer de piel tipo no melanoma presenta la incidencia más elevada (>100 veces) en este tipo de pacientes, pero aún no se ha asociado ningún agente etiológico que justifique esta situación.

A lo largo de esta tesis, se han secuenciado (secuenciación masiva de nueva generación) diferentes lesiones de piel con el fin de determinar los organismos presentes en la epidermis. La mayoría de las secuencias virales obtenidas (>95%) correspondieron al virus del papiloma humano (HPV). Tradicionalmente, la detección de virus ha venido realizándose mediante la reacción en cadena de la polimerasa (PCR). En esta tesis se utilizaron múltiples pares de primers y primers degenerados con el objetivo de amplificar un gran número de HPVs, obteniendo una gran sensibilidad. Sin embargo, aquellos genotipos cuyas secuencias no fuesen similares a las secuencias de los primers, pudieron no haberse amplificado, y por tanto, no ser detectados. Para obviar esta limitación se optó por realizar un protocolo no sesgado (WGA), independiente de la secuencia a amplificar, para determinar si había más genotipos de HPV y/o otros virus presentes en las diferentes lesiones de piel.

Esta tesis ha permitido identificar hasta casi 100 secuencias pertenecientes a posibles nuevos tipos de HPV y caracterizar 4 nuevos genotipos (HPV 197, 200, 201 and 202). La mayoría de los HPVs se detectaron en muy pocos pacientes cada uno y en una concentración viral baja (<0.5 copias/célula), a excepción del HPV 197, que fue el genotipo encontrado con mayor frecuencia (presente en el 37,4% de las lesiones). A pesar de que los métodos basados en la PCR fueron más sensibles, el método basado en WGA fue capaz de detectar HPV en 37/40 condilomas, que habían sido previamente clasificados como HPV-negativos tras genotiparse vía PCR. Algunos HPV, como el tipo 197, fueron detectados solo con el protocolo basado en WGA. Por lo tanto, estimamos que son necesarios más métodos no sesgados, imparciales, para descubrir cuáles son los organismos presentes con mayor frecuencia en las lesiones de piel.

El trabajo realizado en esta tesis ha incrementado nuestro conocimiento sobre la gran diversidad genómica del HPV en la piel. Se necesitan más estudios para evaluar cualquier posible asociación de una infección viral con el cáncer, dilucidar los mecanismos para su desarrollo y determinar la dirección de causalidad.

LIST OF PUBLICATIONS

This thesis is based on the following papers:

I. **Arroyo Mühr LS**, Smelov V, Bzhalava D, Eklund C, Hultin E, Dillner J. Next generation sequencing for human papillomavirus genotyping. J Clin Virology 2013;58:437-42.

II. Ekström J, **Arroyo Mühr LS**, Bzhalava D, Söderlund-Strand A, Hultin E, Nordin P, Stenquist B, Paoli J, Forslund O, Dillner J. Diversity of human papillomaviruses in skin lesions. Virology 2013;447:300-11.

III. Bzhalava D, **Arroyo Mühr LS**, Lagheden C, Ekström J, Forslund O, Dillner J, Hultin E. Deep sequencing extends the diversity of human papillomaviruses in human skin. Scientific Reports 2014;4:5807.

IV. **Arroyo Mühr LS**, Hultin E, Bzhalava D, Eklund C, Lagheden C, Ekström J, Johansson H, Forslund O, Dillner J. Human papillomavirus type 197 is commonly present in skin tumors. Int J Cancer 2015;136:2546-55.

V. **Arroyo Mühr LS**, Bzhalava D, Lagheden C, Eklund C, Johansson H, Forslund O, Dillner J, Hultin E. Does human papillomavirus-negative condylomata exist? Virology 2015;485:283-8.

TABLE OF CONTENTS

ABBREVIATIONS

Aa	Amino acid
AK	Actinic keratosis
BCC	Basal cell carcinoma
bp	Base pairs
CI	Confidence interval
CIN	Cervical intraepithelial neoplasia
dNTP	Deoxyribonucleotide triphosphate
dsDNA	Double-stranded DNA
EmPCR	Emulsion PCR
EV	Epidermodysplasia verruciformis
FFPE	Formalin-fixed paraffin-embedded
GAAS	Genome relative Abundance and Average Size
GASiC	Genome Abundance Similarity Correction
GRAMMy	Genome Relative Abundance estimates based on Mixture Model theory
GS	Genome Sequencer
HPV	Human papillomavirus
HR	High risk
HTS	High throughput sequencing
IARC	International Agency for Research on Cancer
KA	Keratoacanthoma
LR	Low risk
MDA	Multiple displacement amplification
MID	Multiplex identifier
NGS	Next generation sequencing
NMSC	Non-melanoma skin cancer
nt	Nucleotide
OLC	Overlap/Layout/Consensus
PCR	Polymerase chain reaction
PGM	Personal Genome Machine
PV	Papillomavirus
RT-PCR	Reverse-transcription PCR
SCC	Squamous cell carcinoma
SIR	Standardized incidence ratio
SNP	Single nucleotide polymorphism
ssDNA	Single-stranded DNA
UV	Ultraviolet
WGA	Whole genome amplification
WHO	World Health Organization

1 INTRODUCTION

1.1 NON-MELANOMA SKIN CANCER

The skin provides protection and receives sensory stimuli from the external environment, being the largest organ in the body. It is composed of three primary layers: epidermis, dermis and hypodermis (Figure 1).

Most skin cancers arise from the epidermis and are named for the type of cells that become malignant. There are three major types of skin carcinoma:

- **Basal cell skin cancer (BCC),** the most frequently occurring form of skin cancer. This carcinoma arises in the skin's basal cells, which compose the deepest layer of the epidermis.

- **Squamous cell skin cancer (SCC),** the second most common form of skin cancer. This tumor is an uncontrolled growth of abnormal cells arising in the squamous cells, which line most of the epidermis´ upper layers.

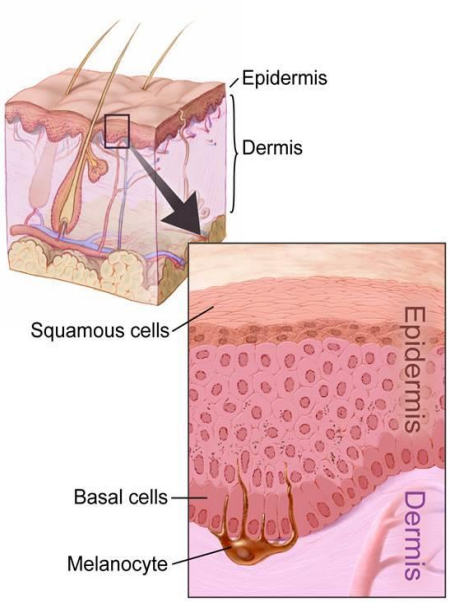

Figure 1: Schematic picture of a cross section of the skin. Image reprinted with permission from the National Cancer Institute, 2016. Illustrator: Don Bliss.

- **Melanoma,** the most dangerous form of skin cancer. This type of cancer originates in the melanocytes, the pigment-producing cells located in the basal layer of the epidermis.

BCC and SCC are generally grouped together as non-melanoma skin cancers (NMSCs) to distinguish them from melanoma, which develops from very different cells and is treated differently, because it is more likely to metastasize.

About 80% of all NMSC are BCCs while SCCs constitute up to 20% [1]. There are a few other rarer types which represent only 1% of skin cancers [1], including keratoacanthomas (KA) - a benign tumor, characterized by a rapid onset followed by spontaneous regression within a few months, Merkel cell carcinomas, cutaneous lymphomas, Kaposi sarcomas and skin adnexal tumors and sarcomas, all of which are classified as non-melanoma skin cancers.

The worldwide incidence of skin cancers has been increasing over the past decades [2-4]. However, it is very difficult to gather overall statistics on NMSC as BCC commonly does not enter the cancer data collection system, since this type of skin cancer is generally treated successfully by dermatologists, and therefore, does not require hospitalization [2]. Consequently, national rates are not available for many countries and worldwide NMSC statistics are often estimates [2]. Most authors report that the average yearly increase in incidence of non-melanoma skin cancer since 1960 is about 3-8%, worldwide [5, 6]. It is estimated that one in every three cancers diagnosed is a skin cancer; between 2 and 3 million non-melanoma skin cancers and 132,000 melanoma skin cancers occur globally each year (http://www.who.int, accessed on 2016-01-15). These data suggest that the prevalence of these tumors will soon equal that of all other cancers combined.

In the case of Sweden, according to regulations by the Swedish National Board of Health and Welfare, all pathology and cytology departments in Sweden must nowadays report all cases of SCC and BCC (SOSFS 2006:15) to the National Swedish Cancer Register. The registration of SCC started already in 1958 whereas registration of BCC started in 2003 [7]. The quality of the infrastructure of the Swedish Cancer Register in terms of completeness, width of information, and reliability of linkage using the personal identification number is internationally recognized. Nordic Registries are known to be very accurate with an overall completeness of over 95% (almost 100% for solid tumors) [8]. Data collected on NMSC in Sweden shows an increase over the past years in accordance with the overall estimated incidence (Figure 2, Table 1).

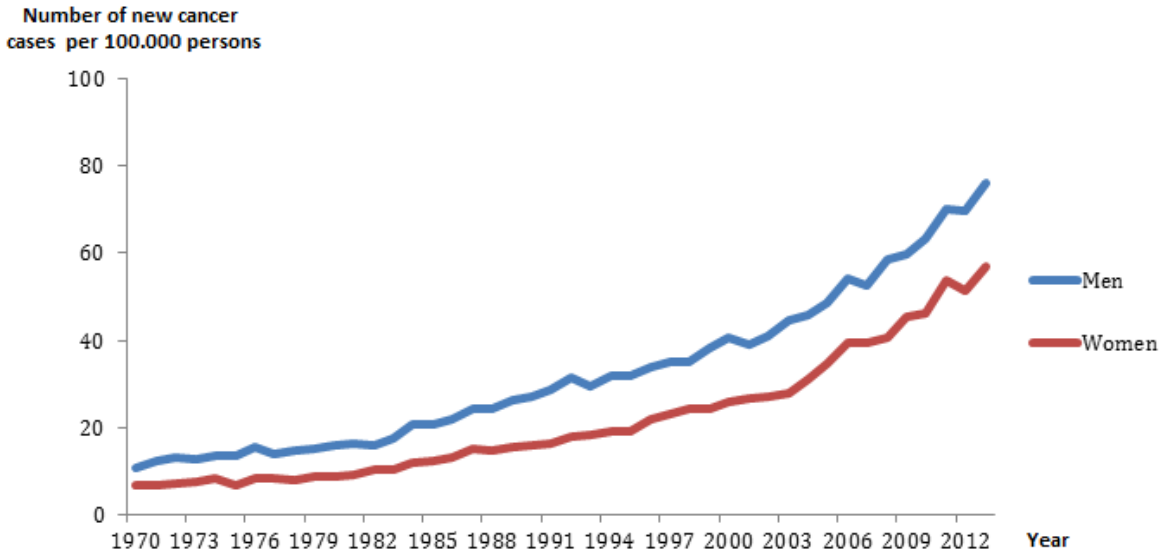

Figure 2: Total number of new skin cancer cases (ICD-7 191. Melanoma excluded) per 100,000 persons (crude rate) in Sweden, 1970-2013. (From the National Board of Health and Welfare´s statistical database, accessed on 2016-01-15).

Year	Men	Women	Total tumors
2004	15632	16138	31770
2005	16369	16554	32923
2006	16489	17209	33698
2007	16710	17919	34629
2008	18165	18395	36560
2009	18103	18813	36916
2010	18333	18605	36938
2011	19783	20052	39835

Table 1: Total number of BCCs by gender, 2004-2011. (From the National Board of Health and Welfare´s statistical database, accessed on 2016-01-15).

The fact that NMSC occurs mainly on sun-exposed sites and that its prevalence can be reduced by sun-protection, provides indirect but crucial evidence for the etiology of ambient solar radiation. While the role of cumulative sun exposure in SCC is well established, the association between sun exposure and BCC seems to rely on intermittent sun exposure and exposure during childhood. Even though we have identified the most

important risk factor and despite growing public awareness of harmful consequences of sun exposure, NMSC incidence continues to increase.

The rising incidence of NMSC might partly be explained by increased patient and physician awareness of the disease, improved coding, as well as an age shift in the population. Furthermore, it might also suggest that other factors might have a role in promoting NMSC in addition to ultraviolet (UV) radiation.

Several complex genotypic, phenotypic and environmental factors contribute to pathogenesis of NMSC. Older age, male sex, fair skin, blond hair, blue eyes, weakened or suppressed immune system [9-12], and a number of inherited genetic skin conditions like epidermodysplasia verruciformis (EV), influence cancer development (http://www.cancer.net, accessed on 2016-01-15).

1.2 INFECTIONS AND NON-MELANOMA SKIN CANCER

The last few decades have led to the realization that a considerable proportion of cancers develop due to infections [13]. Considering infectious agents classified by International Agency for Research on Cancer (IARC) as carcinogenic to humans, 2.1 million (16.4%) of the total 12.7 million new cancer cases that occurred in 2008 in the world were attributable to infections [13].

To date, eight very different viruses have been identified as carcinogenic in humans, including retroviruses (Human T-cell leukaemia virus type I and Human immunodeficiency virus type I), RNA-viruses (Hepatitis C virus), DNA viruses with retroviral features (Hepatitis B virus), and both large double-stranded DNA viruses (Epstein-Barr virus and Kaposi Sarcoma-associated herpesvirus) and small double-stranded viruses (Human papillomavirus, HPV; and Merkel Cell Polyomavirus) [14, 15]. In addition, a bacterium (*Helicobacter pylori*) and some parasites are also clearly implicated in human cancer [13].

All oncogenic infectious agents identified so far have the ability to establish persistent infection in their host. The immune system controls the replication of infectious agents (particularly viruses) and/or expansion of infected cells. Immunosuppression is accompanied by a higher fraction of infection-associated cancers [11, 16-18].

Nordic, including Swedish, registry linkage studies were influential in establishing the large excess risk of cancer among transplant recipients more than 10 years ago [11, 19]. A few cancer types such as brain, breast, corpus uteri, and prostate cancers show no significantly increased incidence in immunosuppressed patients compared to the general population [15, 19]. However, majority of cancers (as well as the overall cancer incidence) is greatly increased among these patients [15, 19].

Most cancer types that are increased among transplant recipients are known to be caused by viruses, e.g. HPV-associated anogenital cancers, Epstein-Barr virus-associated lymphomas, Merkel cell carcinomas and HepatitisB/HepatitisC-associated liver cancers [18], implying that immunosuppression specifically induces an impaired ability to control tumorgenic viruses.

As was pointed out by the 2008 Nobel Laureate Harald zur Hausen, exploration of any further role of infections in cancer is likely to be particularly rewarding if it is focused on the cancer forms that are increased among the immunosuppressed, but that do not have an established microbiological etiology [14]. Non-melanoma carcinoma of the skin (NMSC), including squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), is by far the most highly increased disease in this patient group (about 100-fold increased incidence) [18], but an infectious etiology has not been established.

1.3 POTENTIAL PATHOGENS IN SKIN CANCER

The extreme variety of infectious agents potentially involved in human cancer rules out the possibility of predicting promising candidates. However, human papillomaviruses and bacteria (*Staphylococcus aureus*) have previously been detected in SCC [20-23].

The experiments carried out in this thesis have searched for viruses in NMSC (HPVs as well as other viruses). Metagenomic sequencing found that >95% of the viral sequences present in skin samples belonged to the *Papillomaviridae* family [24] and no other specific virus was commonly detected in most skin cancer specimens. Most studies so far have been carried out after using general polymerase chain reaction (PCR) systems and therefore, they are biased to detect only viruses with sequences of high similarity to the PCR primers used. Viruses that present low similarity to the primer sequences may have remained undetected in previous studies.

1.3.1 HPV

1.3.1.1 Nomenclature and classification

The genus Papillomavirus is a group of small, non-enveloped DNA viruses known since antiquity but first described in the 1930's [25]. The name "Papilloma" comes from the Latin term "papilla" (pustule or nipple) and the Greek suffix "oma" (tumor). Papillomaviruses are identified by the abbreviation PV and one or two letters indicating the host species. For example, human papilloma virus is identified as HPV.

Papillomavirus isolates were traditionally described as "types". The rapid increase in the number of isolates identified demonstrated a need for a taxonomic classification within the family [26]. The first attempt to classify all types relied on the ability of the viruses to infect the squamous epithelium (skin types) or the mucosal epithelium (mucosal types). However, this classification was found to be incorrect due to the possible presence of the same type in both types of epithelium.

HPV classification and nomenclature is based on sequence analysis, as inefficient cell culture systems have limited the possibilities for classification based on biological properties. Both the ability to obtain amplification products based on the PCR technique, and the high stability and conservation of HPV genomes over evolution, support the current classification system based on the differences found in the genome [26],

particularly in the L1 open reading frame, which is the most conserved gene in the genome and encodes for the major capsid protein.

Classifications are as follows (Figure 3, Table 2):

- Genus: different genera within a family share less than 60% nucleotide sequence identity. Currently, human papillomaviruses are divided in five different genera (*alpha*, *beta*, *gamma*, *mu* and *nu*).
- Species: different species within a genus share between 60% and 70% nucleotide identity. There are a total of 49 species, which are designated with a number.
- Genotype (type): genotypes within a species share between 71 and 89% nucleotide homology.

The International Committee on Taxonomy of Viruses is responsible for the papillomavirus nomenclature down to the species [26-28]. In order to establish a potential novel HPV type officially, the whole viral genome must be cloned (more than one plasmid can be used) and sent to the International HPV Reference Center together with the sequence. The International HPV Reference Center will confirm the DNA sequence and assign the submitted clones the novel HPV genotype number, if it is novel.

Today, 205 HPV different genotypes have been completely cloned, sequenced and given an official number at the International HPV Reference Center (http://www.hpvcenter.se, accessed on 2016-01-15). Four previously awarded HPV type numbers (HPV46, HPV55, HPV64 and HPV79) were withdrawn (mostly due to re-classification as subtypes of other HPV types). There are therefore 201 different HPV types established today.

The number of HPV types is continuously growing. The exact number of putative novel types is difficult to ascertain, mostly due to the possibility of different non-overlapping partial sequences representing the same virus type.

It is estimated that at least 400 different HPV types exist [29]. In the HPV center, 360 putatively novel HPV types have been already discovered [30-33]. The *gamma* genus is rapidly growing with now up to 81completely HPV established types, surpassing *alpha* and *beta* genera, with 65 and 51 types, respectively (Figure 3, Table 2).

During the last five-year period, 77% of all new HPVs deposited in the International Reference Center belonged to the *gamma* genus. Surprisingly, *mu* and *nu* genera have almost not increased in number. Methods that are independent of sequence information have not revealed any additional members. An exception is the recently discovered HPV 204, isolated from the anal canal.

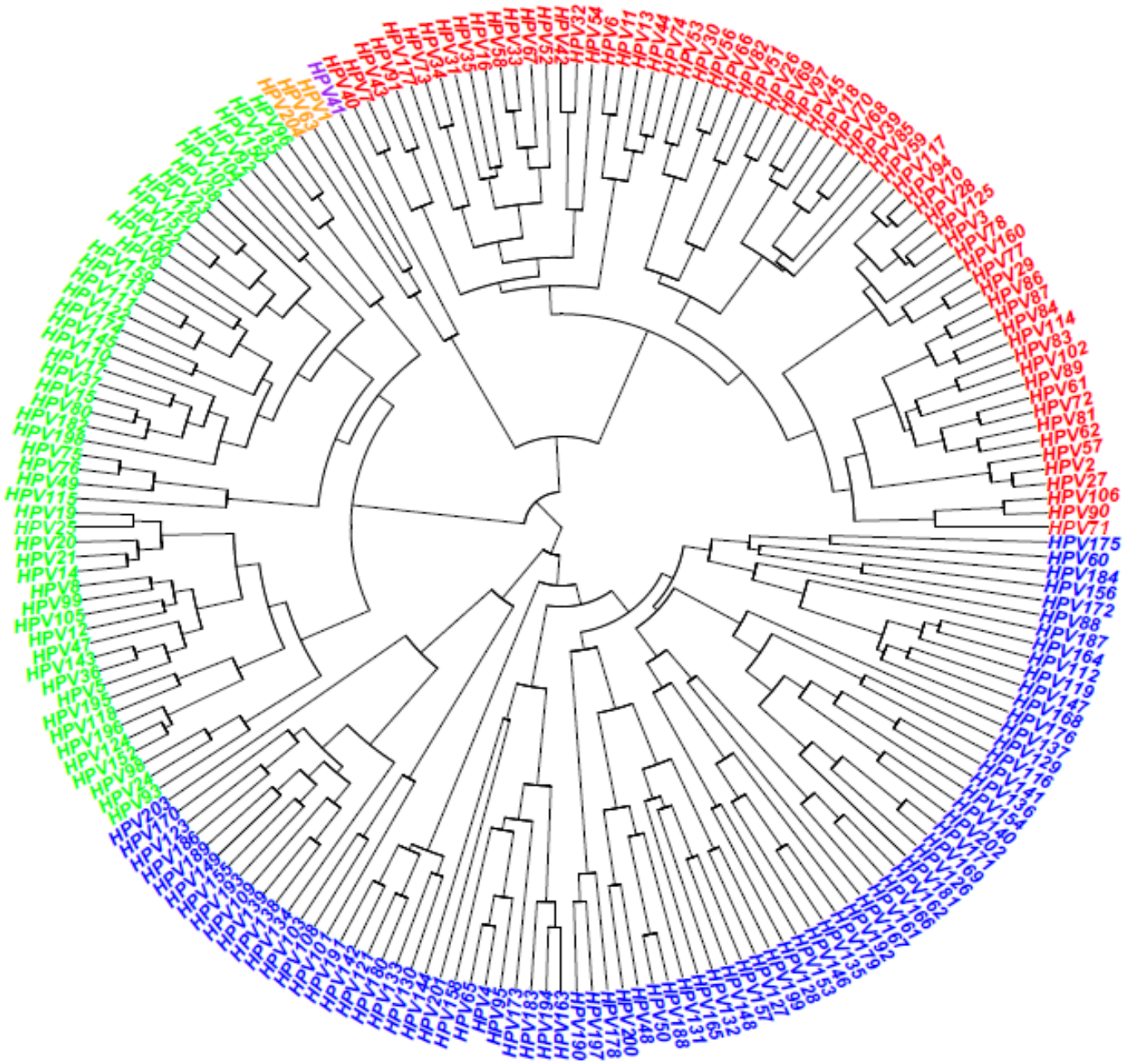

Figure 3: Phylogenetic tree of 204 HPV types. *Alpha*, *beta*, *gamma*, *mu* and *nu* papillomaviruses are presented in red, green, blue, orange and purple colors, respectively. The phylogenetic tree is based on the L1 part of the genome.

Genus	Species	First HPV type	Other HPV types	Date
Alpha	Alpha-1	HPV32	42	1986-1987
	Alpha-2	HPV3	10, 28, 29, 77, 78, 94, 117, 125, 160	1984-2009
	Alpha-3	HPV61	62, 72, 81, 83, 84, 86, 87, 89, 102, 114	1989-2008
	Alpha-4	HPV2	27, 57	1984-1989
	Alpha-5	HPV26	51, 69, 82	1985-1997
	Alpha-6	HPV30	53, 56, 66	1981-1987
	Alpha-7	HPV18	39, 45, 59, 68, 70, 85, 97	1981-2004
	Alpha-8	HPV7	40, 43, 91	1984-2001
	Alpha-9	HPV16	31, 33, 35, 52, 58, 67	1984-1989
	Alpha-10	HPV6	11, 13, 44, 74	1984-1993
	Alpha-11	HPV34	73, 177	1985-2013
	Alpha-13	HPV54		1987
	Alpha-14	HPV71	90, 106	1991-2004
Beta	Beta-1	HPV5	8, 12, 14, 19, 20, 21, 24, 25, 36, 47, 93, 98, 99, 105, 118, 124, 143, 152, 195, 196	1984 - 2014
	Beta-2	HPV9	15, 17, 22, 23, 37, 38, 80, 100, 104, 107, 110, 111, 113, 120, 122, 145, 151, 159, 175, 182, 198	1984-2014
	Beta-3	HPV49	75, 76, 115	1987-2008
	Beta-4	HPV92		2001
	Beta-5	HPV96	150, 185	2002-2013
Gamma	Gamma-1	HPV4	65, 95, 95, 158, 173, 205	1984-2015
	Gamma-2	HPV48	200	1987-2014
	Gamma-3	HPV50	188	19872013
	Gamma-4	HPV60		1989
	Gamma-5	HPV88		2001
	Gamma-6	HPV101	103, 108	2004-2006
	Gamma-7	HPV109	123, 134, 138, 139, 149, 155, 170, 186, 189, 193	2007-2014
	Gamma-8	HPV112	119, 147, 164, 168, 176	2007-2013
	Gamma-9	HPV116	129	2009
	Gamma-10	HPV121	130, 133, 142, 180, 191	2009-2013
	Gamma-11	HPV126	136, 140, 141, 154, 169, 171, 181, 202	2010-2014
	Gamma-12	HPV127	132, 148, 157, 165, 199	2009-2014
	Gamma-13	HPV128	153	2009-2011
	Gamma-14	HPV131		2009
	Gamma-15	HPV135	146, 179, 192	2009-2013
	Gamma-16	HPV137		2009
	Gamma-17	HPV144		2010
	Gamma-18	HPV156		2011
	Gamma-19	HPV161	162, 166	2012
	Gamma-20	HPV163	183, 194	2012-2014
	Gamma-21	HPV167		2012
	Gamma-22[a]	HPV172		2012
	Gamma-23[a]	HPV175		2013
	Gamma-24[a]	HPV178	190, 197	2013-2014

Genus	Species	First HPV type	Other HPV types	Date
Gamma	Gamma-25[a]	HPV184		2013
	Gamma-26[a]	HPV187		2013
	Gamma-27[a]	HPV201		2014
	Unclassified	HPV203		2014
Mu	Mu-1	HPV1		1984
	Mu-2	HPV63		1991
	Unclassified	HPV204		2014
Nu	Nu-1	HPV41		1987

Table 2: Established HPV types, stratified by species and genera. Date refers to the period when HPV types were officially assigned with an established number by the International HPV Reference Center. [a]: Species assignments are tentative and not official, but recommended to the papilloma virus working group of ICTV. Modified table from www.hpvcenter.se, accessed on 2016-01-15.

1.3.1.2 HPV in Skin Cancer

The papillomavirus life cycle is tightly linked to the differentiation process of the infected epithelium. Papillomaviruses initially infect basal epithelial cells, which constitute the only cell layer in an epithelium that actively divides. The mechanisms by which HPV induces neoplastic transformation are probably various, and in fact, in vitro models demonstrate only a weak potential. It (neoplastic transformation) is attributed in a large part to the actions of the HPV *E6* and *E7* oncogenes [34-36].

These oncoproteins inactivate tumor suppressor genes that operate at key cell cycle checkpoints. E6 interferes with p53, leading to genomic instability and blocking of apoptosis, allowing cells with damaged DNA to replicate rather than self-destruct while E7 inactivates retinoblastoma signaling, leading to induction of DNA synthesis in keratinocytes that would otherwise be terminally differentiated and non-replicating [37].

Interestingly, SCCs derived from mice with a deletion of the retinoblastoma protein or the *p53* gene only in skin, exhibit similar molecular signatures to that of HPV-induced tumors, suggesting a role of HPV in the carcinogenesis of SCC [38]. Thus, when E6 and E7 act synergistically, not only do they promote inhibition of apoptosis and dysregulation

of the cell cycle leading to abnormal cell growth, but also induce cellular genomic instability contributing to carcinogenesis. However, the effect by itself is not enough to transform cells [39].

UV exposure is an important cofactor in HPV carcinogenesis. It may be then, that the contribution of HPV infection to cancer is via the anti-apoptotic effect in UV-damaged keratinocytes, which would have otherwise progressed to senescence and disintegration. This inhibition of apoptosis probably results in persistent viral infection and hence the accumulation of further DNA mutations, putatively leading to immortalized cells. Unrepaired DNA damage has been observed in UVB-irradiated cells expressing the E6 protein, and inactivation of the retinoblastoma protein with HPV 16 E7 has resulted in significant inhibition of the ability to recover mRNA synthesis and increased levels of apoptosis following UV radiation [40, 41].

An association between HPV and NMSC has been found among patients with epidermodysplasia verruciformis (EV), a rare hereditary immunosuppressive disease [42]. EV patients develop skin lesions in early infanthood and present eruptions of wart-like lesions which are refractory to conventional wart treatment and progress to SCC at sun-exposed sites of the skin [43]. The persistence of HPV infection in EV has been suggested to be due to the inability of the patient's immune system to reject HPV-infected keratinocytes by a still unknown immunogenetic defect and is probably also influenced by environmental factors, particularly ultraviolet radiation [44]. The HPV types found in patients with EV are referred to as EV-HPV types, and include, among others, HPV types 5, 8, 9, 12, 14, 15, 17, and 19–25 [45, 46]. HPV 5 and 8 are the most prevalent types [47].

In contrast to cervical cancer where HPV genotypes 16 and 18 have been established as the most prevalent genotypes which cause this disease (70% of cancer cases) and in contrast to patients suffering from EV where HPV 5 and 8 are high-risk genotypes for skin cancer, the HPV types found in skin cancers of the general population have varied depending on which PCR-system was used [48, 49]. It is common to detect multiple genotypes in a single specimen [24, 50].

Metagenomic sequencing has revealed that >95% of the viral sequences present in NMSC samples belong to the *Papillomaviridae* family, mostly to the *beta* and *gamma* genera

[24] but so far, only one study, that is included in this thesis, has detected a particular genotype (HPV 197) with high frequency (37.4%) [30] .

Infections of HPV in skin are very common, but because of the diversity of HPV types, there does not appear to exist any single virus that is widely spread. Acquisition appears to occur already shortly after birth [51-54]. A broad spectrum of cutaneous HPV is commonly detected both on healthy skin [55, 56], in plucked eyebrow samples [57-59] as well as in different skin diseases such as SCC, BCC, actinic keratosis (AK) – a precursor lesion for SCC – and in KAs, in both immunocompetent and immunosuppressed patients [60-63].

It has to be highlighted that detection of an HPV type in skin tumors does not necessarily mean that an HPV infection has been detected, as it may merely be a viral contamination of the skin surface. Forslund et al. demonstrated that cleansing of the skin by simple tape stripping before sampling, strongly reduces the proportion of HPV positive samples [62]. Prevalence dropped from 69% in swabs from top of SCCs, BCCs and AK lesions, to 12% in the corresponding biopsies, after cleansing the skin surface.

When skin biopsies from NMSC only contain low viral loads (<1 copy/cell) [22, 64-66], it is debatable whether such low viral copy numbers are biologically relevant to tumor initiation and maintenance.

1.4 VIRAL DETECTION IN SKIN CANCER

When analyzing human skin specimens, one has to take into account that besides human DNA and RNA, human skin harbors various physiological populations of microorganisms, including commensal or symbiotic bacteria, fungi, parasites and viruses, overall known as the skin microbiota.

Sequencing studies reveal that viruses represent < 1% of the total genomic material in skin [31] and therefore, detection of any virus by NGS, formerly required performing some type of viral enrichment or amplification first. Viral enrichment can be achieved by

performing: low-speed centrifugation and/or filtration to remove bacterial and host cells, nuclease treatment to digest nucleic acids that are not protected with virions [67], separation of long chromosomal DNA from shorter DNA [31], high-speed gradient centrifugation [68] or targeted sequence capture [69-71]. Each of these procedures may bias against detection of some viruses and result in decreased assay sensitivity as a result of loss of viral nucleic acids. Novel protocols designed to overcome these problems look very promising. ViroCap for instance, is a viral targeted sequence capture panel, with multiple probes designed to capture most viral species that infect vertebrates (337 viral species). This test is also capable to capture those viruses that share up to 58% variation from the reference viruses used to select capture probes [72].

The publications included in this thesis analyze only DNA virus (no RNA) and thus, amplification techniques and detection methods will focus on DNA. RNA viral presence will be discussed in the section "Concluding remarks and future perspectives".

1.4.1 Amplification techniques

1.4.1.1 PCR using "general" or "degenerated" primers

The relative ease and economic accessibility of the PCR technique made it one of the most widely used techniques in clinical diagnostics. Several general primer PCR systems targeting the L1 gene (FAP, CUT, PGMY, MGP) [33, 48, 73, 74] can amplify a broad range of HPV types. However, efficiency of any PCR based amplification depends on the number, position, and stability of mismatches between the primers and the template. The sequence to be amplified must be previously known and thus, amplification is biased to detect only sequences of high similarity to the primers used. HPV types with low similarity to the primer sequences will remain undetected.

As an example, Forslund et al., designed PCR FAP primers from two relatively conserved regions of the L1 open reading frame taking into consideration most known genome sequences, from HPV 1 to HPV 80 [49]. FAP primers could detect 65/75 (86.7%)

different HPV types but were not able to amplify HPV types 1, 2, 35, 41, 44, 55, 63, 66, 71 or 74. The failure in detecting HPV1, 41 and 63 is of particular importance as these types are the only genotypes that form the genera *mu* and *nu* (recently, HPV 204 was also classified as a *mu* papillomavirus (hpvcenter.se)). Other putative novel genotypes phylogenetically close to these HPV types (belonging to these genera) might also remain undetected.

Currently, there are 205 different HPV types recognized. Attempts to degenerate FAP primers more, in order to amplify a broader number of HPV types, as well as multiplexing specific HPV primers in the PCR reaction, might improve HPV detection, but will at the same time reduce specificity in the reaction (e.g. human DNA binding).

1.4.1.2 Unbiased approach: Whole genome amplification

By performing whole genome amplification (WGA), all the DNA present in a sample will be amplified without requiring previous knowledge of the DNA sequence. There are different methods that have been developed for high-fidelity WGA.

Initial PCR-based WGA techniques were based on the use of degenerate/semi-degenerate primers [75] and primer extension PCR [76]. These techniques used Taq polymerase and consequently, limited the amplicon fragment length to 3 kb and introduced errors in the sequence due to the lack of the enzyme´s 3´- 5´ exonuclease activity. Furthermore, they exhibited incomplete genome coverage and amplification bias [75-77] and therefore, were substituted by non-PCR based methods.

Multiple displacement amplification (MDA) is today a gold standard method for non-PCR based amplification techniques. It was first developed for rolling circle amplification [78] where the reaction starts by annealing random hexamer primers to the DNA template and DNA synthesis is carried out at constant temperature (Figure 4).

The DNA polymerase used in MDA originates from bacteriophage phi29 [79, 80] and has a 3´- 5´ proofreading activity, resulting in a low intrinsic error rate and about 1000-fold

less accumulation of mutations compared to PCR techniques using Taq DNA polymerase [81, 82]. Moreover, Phi29 possesses a strong strand displacement activity, being able to solve secondary structures as hair pin loops, thereby preventing slipping, stopping and dissociation of the polymerase during amplification. Average product length can be greater than 10 kb [83].

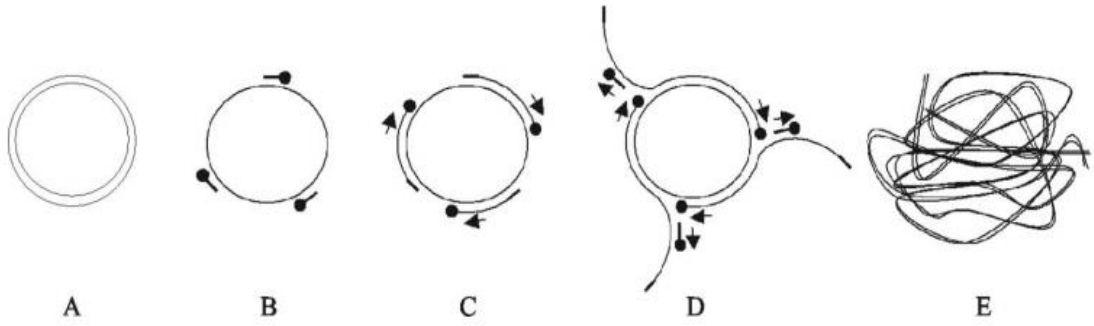

Figure 4: Overview of the whole genome amplification with focus on double stranded DNA (A). Random hexamer primers bind to single stranded DNA and amplification starts (B, C). When the polymerase reaches a downstream primer, the strand is displaced and new primers can anneal to the displaced product (D). The end product is double stranded repeated copies of the DNA in the sample (E). Figure adapted from Rector et al., A sequence-independent strategy for detection and cloning of circular DNA virus genomes using multiply primed rolling-circle amplification, in Journal of Virology, 2004; 78:4993-8, with permission from American Society for Microbiology.

MDA provides an effective and easy means of amplifying minimal quantities of DNA and is the least biased WGA method reported [83, 84]. However, there are also biases associated with this technology. Chimera formation, preferential amplification of circular single stranded DNA (ssDNA) and non-uniform amplification of linear genomes have been documented [85, 86].

The predominant mechanism for chimera formation is now elucidated [87]. More than 85% of chimeras are due to inverted sequences occurring when strand displacement takes place in the reaction. 3´-termini can be displaced and might reanneal at randomly occurring complementary segments on nearby 5´-strands, resulting in the joining of two

sequences in inverted orientation with an intervening deletion. Bioinformatic analysis might be able to find these sequences and remove them for further analysis.

MDA has not shown the ability to accurately estimate the amount of viral populations present [88], most probably due to the preferential amplification of particular genomic regions during initial MDA priming events [89, 90]. Several investigators have proposed that pooling several independent MDA reactions run on a single sample of template DNA minimizes representational bias in shotgun metagenome sequence libraries [91-96]. However, this assumption has not been thoroughly tested [97].

There is no reported method that overcomes all amplification bias in MDA products. Hence, the effective use of MDA in any application depends on the user´s needs. E.g.: if the user is only interested in ssDNA sequences, skipping the denaturation step in the amplification will completely bias the ssDNA amplification, as primers will not be able to bind to non-denatured dsDNA [98].

Even though MDA was traditionally used to amplify circular DNA [96], it can be used to amplify linear DNA [83, 100]. In Paper III, we quantified the amount of amplification of both human DNA and HPV DNA by adding 20 copies/μL of HPV 16 plasmid to samples of human placental DNA at 1 ng/μL. We amplified these samples in the same manner as for the clinical samples and quantified the amounts using real-time PCR for beta-globin and for HPV 16, respectively. Human DNA was found to be amplified 26-fold, whereas HPV 16 DNA was amplified 679-fold. Thus, although the WGA will have made it easier to detect the circular HPV genomes, it is unlikely that we would have missed linear and/or large dsDNA viruses unless they were present in only small amounts.

1.4.2 Detection techniques

Detection of amplified products is usually performed by hybridization of amplicons to type-specific probes coupled to fluorescent beads [101, 102] or by product sequencing [20, 21, 48, 103].

1.4.2.1 Hybridization to type-specific probes

Detection of amplicons using probes requires prior knowledge of the DNA sequence query in order to design specific probes. There is a large variety of molecular assays based on hybridization used for detection of different microorganisms [104, 105]. Some of the hybridization methods for HPV DNA typing include the Hybrid Capture II test (Digene Corporation, Gaithersburg, MD), the LINEAR ARRAY® HPV genotyping test (Roche Molecular Systems, Alameda, CA), the INNO LiPA® HPV genotyping test (Fujirebio, Gent, Belgium), and the Cervista® HPV (Hologic Inc., Marlborough, MA).

Accurate and internationally comparable DNA detection and typing methodology is an essential component both for research and for diagnosis. The World Health Organization (WHO) started a WHO Global HPV Laboratory Network (LabNet) in 2006 to support the world-wide development and implementation of HPV vaccines through improved laboratory standardization and quality assurance of HPV testing and typing methods (Technical Report on the Global HPV LabNet 2013 HPV DNA Genotyping Proficiency Panel). The results show a yearly increase in proficiency (sensitivity and specificity) of HPV typing assays when routinely used in laboratories worldwide [106, 107].

Despite a high sensitivity and specificity of these methods, their major limitation is the potential for error in HPV typing because of probe cross-hybridization when cross-reactivity of one probe to several target groups occurs (false positivity). Furthermore, sequences with low similarity to the probes or sequences that present small variations and point mutations might not be detected (false negativity). Therefore, these methods are not valid when searching for novel viruses and/or sequences.

A recent study developed a comprehensive viral targeted sequence capture panel using hybridization probes, that were able to assess all viruses known to infect vertebrate cells (excluding human endogenous retroviruses), as well as to detect divergent viruses [72]. The main advantages of this design are: complete viral genomes targeting, detection of viruses that show up to 58% variation from the reference virus used to select capture probes and, the possibility to isolate the capture sequences in order to sequence them afterwards [72].

1.4.2.2 High throughput sequencing

With the establishment of the high throughput sequencing (HTS) technology (also called next generation sequencing (NGS) or deep sequencing), there is no need to test specimens for predefined microorganisms by PCR. Sequencing the DNA will determine which organisms are present by analyzing the sequence data. Furthermore, the entire sequence is obtained and this enables, both discovering novel viruses as well as finding small variations and point mutations within the previously known viruses.

1.5 HIGH THROUGHPUT SEQUENCING INSTRUMENTS

During the past decade, there has been a dramatic evolution of next generation sequencing technologies: 454 GS (Roche), SOLiD (ABI), Ion Torrent (Life Technologies) and Genome Analyzer System (Illumina).

These recent technological advances have revolutionized the study of genomics and molecular biology. NGS allows us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing. Using capillary electrophoresis-based Sanger sequencing, the Human Genome Project took over 10 years and costed nearly $3 billion. Nowadays, sequencing a human genome can be done in a variety of bench-top NGS instruments in a couple of days and at very limited costs.

In this thesis, we have used 2 different sequencing technologies. We started sequencing with 454 GS Junior (Roche) but after about a year, quicker, cheaper and deeper platforms came on the market such as the Genome Analyzer System from Illumina (MiSeq, HiSeq and NextSeq platforms). A comparison of some of the platforms´ specifications used in this thesis is shown in Table 3.

Method	Sanger-CE	454 GS Junior	MiSeq	HiSeq	NextSeq
Year	1980's	2000's	2010's	2010's	2014
Output	<100 kb	35 Mb	1500 Mb	600 Gb	120 Gb
Read length	500 bp	400 bp	300x2 bp*	100x2 bp*	150x2 bp*
Run time	7 h	8 h	56 h	11 days	29 h
$ per Mb	2400	31	0.5	0.05	0,45

Table 3: Next generation sequencing platforms and their specifications. *Pair-end sequencing.

1.5.1 454 GS technology (Roche)

The 454 Life Sciences (454; Branford, CT, USA; now Roche, Basel) sequencing platform (the 454 Sequencer) was the first next-generation technology to reach the market. It was first commercially introduced in 2004. It dramatically increased the volume of sequencing conducted by research groups and expanded the range of problems that can be addressed by the direct readouts of DNA sequence.

It was the first technology to sequence and assemble bacterial genomes *de novo* [108] and the first non-Sanger technology to sequence an individual human [109]. Other notable studies conducted by 454 included work as diverse as uncovering the potential cause of the disappearance of the honeybee [110], revealing the complexity of rearrangements between individual human genomes [111], providing new approaches to understand infectious diseases [112] - such as the mechanism of resistance to the drug R207910 in *Mycobacterium tuberculosis* [113] - and sequencing the first million base pairs of a Neanderthal [114-116].

The GS Junior (454) technology is based on emulsion-based amplification and pyrosequencing. Its workflow is comprised of three main steps: generation of a single-stranded template DNA library, emulsion-based clonal amplification of the library, and pyrosequencing (Figure 5).

1.5.1.1 Generation of a template DNA library

The DNA is isolated and fragmented by nebulization. DNA fragments are polished and end repaired to create 3'Adenine ends to allow TA ligation of multiplex identifiers (MIDs) adaptors. These adaptors are very important, as they both allow sample quantification, and permit the user to multiplex sequencing, as each DNA fragment can be ligated to a different MID.

Once the adaptor is ligated, a size selection step is performed to get rid of adaptors that were not ligated to the DNA fragments and samples are purified. Before continuing to the next step, it is important to assess the library quality and quantity in order to obtain the optimal number of molecules of library DNA for emulsion PCR.

1.5.1.2 Emulsion-based clonal amplification of the library

Pools of DNA libraries with different MIDs are combined with capture beads and then amplified by emulsion PCR. The beads are captured by droplets of a PCR reaction mixture in oil emulsion, and PCR amplification occurs in each droplet. This results in each bead carrying ten million copies of a unique DNA template.

The emulsion is then broken, the bead-attached DNAs are denatured and the beads are deposited into wells of a fibre-optic slide, the PicoTiterPlate^{TM}, containing hundreds of thousands of wells, wide and deep enough to contain only one bead.

1.5.1.3 Pyrosequencing

Pyrosequencing is a method of DNA sequencing based on the "sequencing by synthesis" principle. It relies on the detection of pyrophosphate release on nucleotide incorporation. The desired DNA sequence can be determined by light emitted upon incorporation of the

next complementary nucleotide. Only one out of four of the possible A/T/C/G nucleotides are added and available at a time, thus only one letter can be incorporated on the single stranded template (which is the sequence to be determined) and light is emitted (once per nucleotide incorporation). The previous nucleotide letter is degraded before the next nucleotide letter is added for synthesis. The intensity of the light determines if there is more than one of these nucleotides in a row. This process is repeated with each of the four letters until the DNA sequence of the single stranded template is determined.

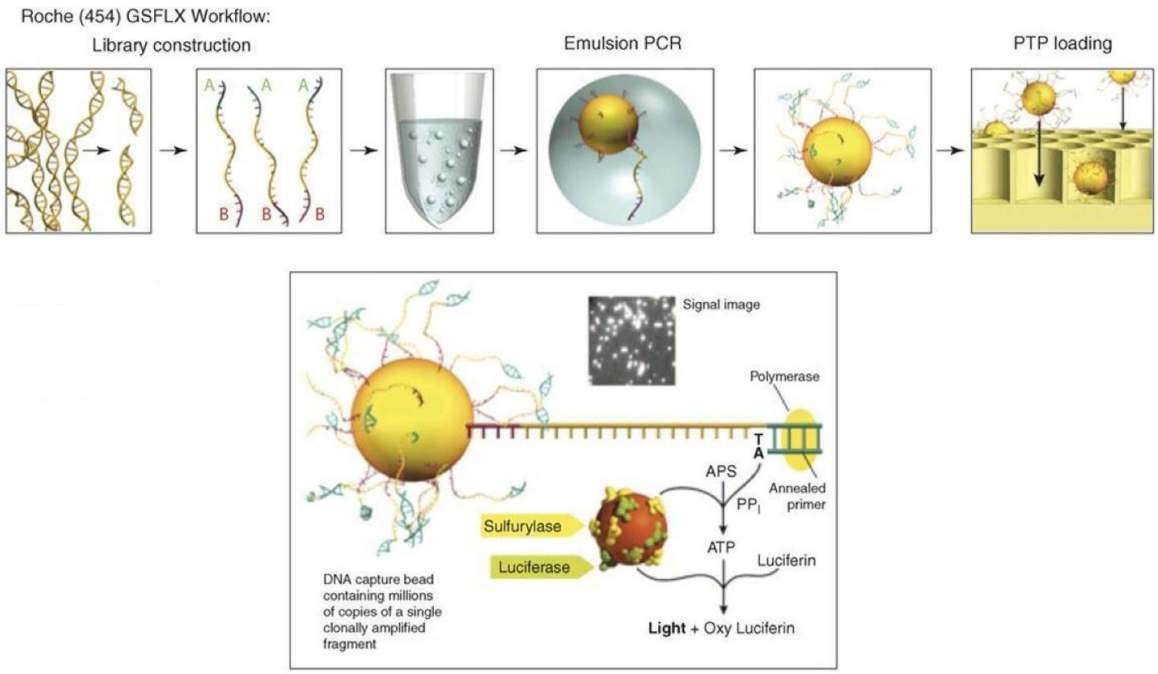

Figure 5: Schematic overview of high-throughput sequencing using 454 GS (Roche). Image reprinted from Mardis et al., The impact of next-generation sequencing technology on genetics, in Trends in Genetics, 2008; 24:133-41, with permission from Elsevier.

1.5.2 Genome Analyzer System technology (Illumina)

The Genome Analyzer System technology was first developed by Solexa Inc., which launched the first Genome Analyzer in 2006. This platform enabled scientists to sequence up to 1 Gb of data in a single run. In early 2007, Solexa was bought by Illumina Inc., and it now supports a broad range of applications. Agrigenomics, cancer, forensics, complex

diseases, drug development, and microbial genomics as well as reproductive and genetic health are the most common fields where this platform has been used. With the added depth of sequencing, it enables: the identification of low-abundance genomes or those exhibiting modest expression differences between samples, discovering all types of genetic variations (SNPs, rearrangements, copy number variants, insertions, and deletions) [117-119], characterization of new bacterial isolates [120-122] and/or new variants that cause diseases, profiling DNA methylation status across the entire genome [123-125], defining somatic variations in cancer [124], and characterizing complex RNA populations for new genes and transcript structures [126, 127], among other utilities.

Illumina's sequencing technology is based on cluster generation and sequencing by synthesis, tracking the addition of labeled nucleotides as the DNA template is copied in a massively parallel fashion. Its workflow follows three main steps: generation of a single-stranded template DNA library, cluster generation, and sequencing by synthesis (Figure 6).

1.5.2.1 Generation of a template DNA library

The DNA sample is prepared into a “sequencing library” by using either an enzymatic fragmentation followed by adding unique adapter sequences (forward and reverse oligomucleotides) to the sample, or an engineered transposome that simultaneously fragments and tags ("tagment") input DNA in the process.

After tagmentation (or fragmentation and adapters ligation), a limited-cycle PCR reaction uses these adapter sequences to amplify DNA fragments. This PCR reaction also adds index sequences on both ends of the DNA fragments, enabling dual-indexed sequencing of pooled libraries on any Illumina Sequencing System.

1.5.2.2 Cluster generation

The sequencing libraries are denatured and attached to a lawn of single stranded oligonucleotides immobilized on a flow cell surface. These olignonucleotides correspond to the complementary sequences of the adapters ligated during the library preparation step. Subsequently, each end of every library molecule matches one of the two primers on the glass surface.

Once the sequencing templates are attached, amplification in a bridge-fashion occurs. The free/distal end of the DNA template loops over to hybridize the complementary surface primer (oligonucleotide). A DNA polymerase copies the templates from the hybridized oligonucleotides forming dsDNA bridges, which are denatured and result in two single strands that serve as a new template, and will then loop over and hybridize again.

Priming will continue as the distal end of a ligated fragment bends over to a complementary oligo on the surface of the flow cell. Repeated denaturation and extension will result in millions of surface-bound colonies, each of them containing approximately one million copies of each template (the cluster).

1.5.2.3 Sequencing by synthesis

The Illumina sequencing method uses fluorescently-tagged dNTPs containing a terminator which blocks further polymerization. During each sequencing cycle, all 4 labelled nucleotides are added to compete for addition (thus, minimizing incorporation bias) to the DNA strand. Only one nucleotide can be incorporated based on the sequence of the template.

The included base behaves as a terminator for the polymerization. After each cycle, the fluorescent dye is imaged to identify the nucleotide and the terminator is cleaved to allow the incorporation of the next base. This process is repeated until the DNA sequence of the single stranded template is determined.

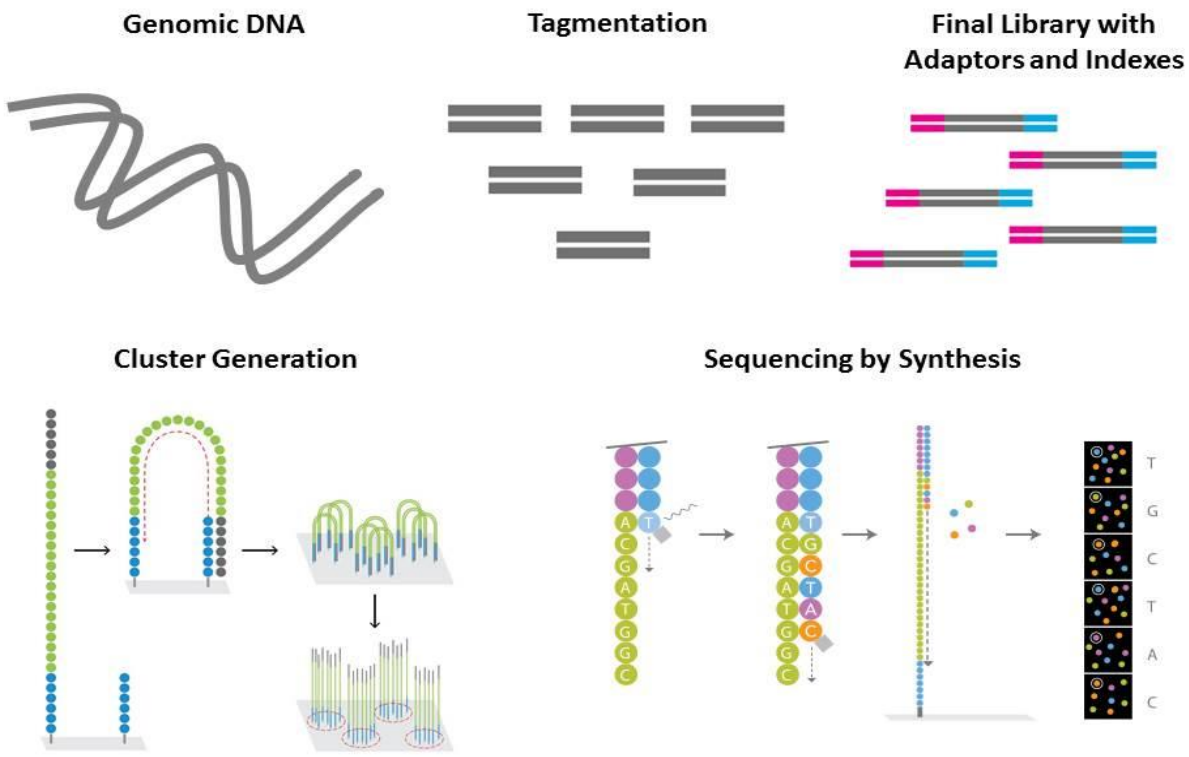

Figure 6: Schematic overview of high-throughput sequencing using Illumina technology.

1.6 HIGH THROUGHPUT SEQUENCING DATA ANALYSIS

As even just one sequencing run provides enormous amounts of data, analysis and interpretation are challenging.

Sequencing analysis usually starts by performing quality checking [128] (Figure 7). Quality checking is an important and effective measure for determining the quality of sample libraries, and it also serves to indicate whether the sequencing succeeded or failed. Bases are checked according to their Phred quality scores [128]. Phred quality scores are logarithmically related to the base-calling error probabilities. For example, a Phred quality score of 10 corresponds to a base calling accuracy of 90% (10 errors per 100bp), while a quality score of 20 equals to a base calling accuracy of 99% (1 error per 100bp) [128].

Specific quality filtering conditions can be adapted for different downstream analyses [129].

To obtain a dataset that contains reads of interest, e.g. the virus-related reads for viral metagenomics, sequences that are not a target of the investigation (e.g. human and bacterial reads) need to be filtered out on the bioinformatics level (Figure 7). This will further speed up the downstream analysis and decrease the risk of mis- assemblies [129].

NGS sequences from human samples subjected to WGA typically contain more than 60 percent human-related sequences. Human- and bacteria-related reads are the most commonly obtained reads, followed by sequences classified as "other" and "unknown" [31, 130] (Table 4). Viral reads usually represent less than 0.5% of sequencing data and enrichment for viral particles by ultracentrifugation has not been shown to be helpful in the analysis of biopsies or skin swabs [130].

Negative control samples (PCR grade water) might also contain bacterial sequences and sequences classified as "other" and "unknown" [130]. Water controls have so far been found to be uniformly negative for viral sequences [130].

These background sequences might be present due to the background reactivity of the Phi29 polymerase reaction [131] or represent environmental contamination and therefore, it is imperative that all metagenomic sequencing projects include sequencing of negative control samples [130].

Once human and bacterial sequences are filtered out, data normalization is performed in order to decrease sample variation and discard redundant data such as duplicated reads (Figure 7). NGS technologies can produce duplicated reads due to errors in PCR amplification and/or sequencing [132, 133] and these reads might introduce an overestimation of the species abundance. Duplicated reads may also include natural duplicates that by chance originate from the same genomic position [132, 133]. Highly abundant species have a higher chance of natural duplicates [133] and their removal might introduce bias towards underestimation of abundances [132].

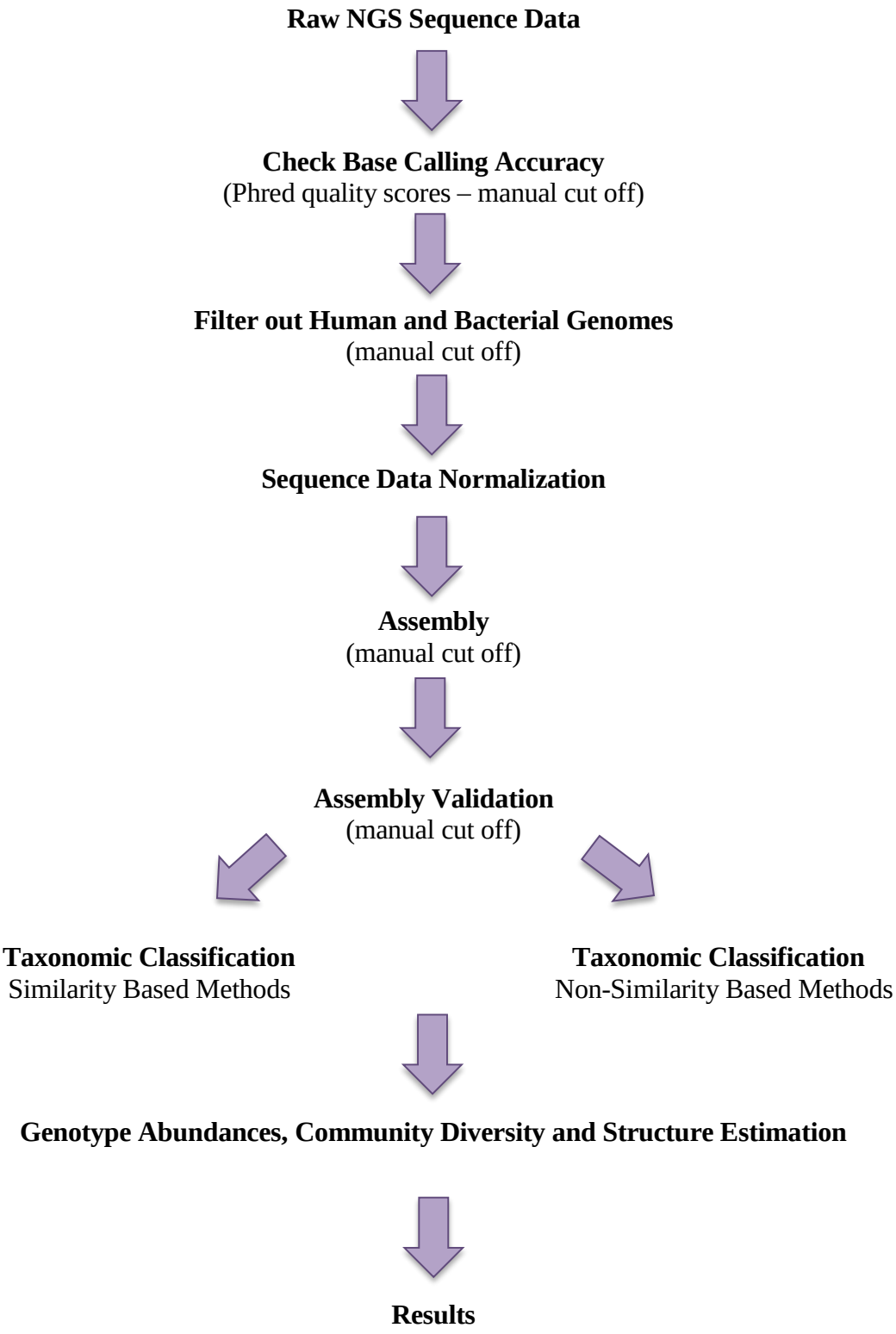

Figure 7: Bioinformatics pipeline to analyze high-throughput sequencing data for viral metagenomics.

	FFPE Biopsies	Biopsies	Skin swabs		Water
Sequencing platform	GSFLX	GSFLX	GSFLX	PGM 400	GSFLX
Human	37,3	99,8	69,1	76,3	2,8
Bacteria	21,3	0,1	24,2	18,3	52,2
Virus	0,2	0	0,3	0,3	0
Other	10,2	0	2,2	1	15,5
Unknown	30,9	0	4,2	4,1	29,5

Table 4: Typical taxonomic assignment of NGS reads (%). Summary of results in previous studies, using different types of biospecimens, pre-treatments and NGS platforms. FFPE: Formalin Fixed Paraffin Embedded. Adapted from Bzhalava et al., Unbiased approach for virus detection in skin lesions, in PloS One, 2013; 8:e65953, with permission from the Creative Commons Attribution License.

Sequence datasets are usually normalized using a digital normalization algorithm (http://ged.msu.edu/papers/2012-diginorm), which substantially reduces data size and computational resources for *de novo* assembly. NGS technologies produce billions of short reads from random locations in the genome by oversampling it and assembling those reads is the next step performed in the bioinformatics analysis (Figure 7).

Assembly algorithms, in a process called *de novo* assembly, reconstruct original genomes which are present in the sample by merging short genomic fragments into longer contiguous sequences ("contigs"). There are two main types of *de novo* assembly programs: Overlap/Layout/Consensus assemblers, most widely applied to the longer reads and *de Bruijn Graph* assemblers, applied to the shorter reads. To validate assembly results, several assembly algorithms might be used, as well as re-mapping all singletons reads to assembled contigs [31, 134].

The possibility always exists that assembly algorithms may construct erroneous "chimeric" sequences by assembling sequences from different organisms or species. This problem may be particularly relevant for viral metagenomics where the biospecimens may contain a multitude of related viral sequences. For HPVs, we developed an algorithm to identify possible "chimeric" HPV sequences [135]. It is based on the assumption that an HPV genome should have similar degree of identity to the most closely related HPV type over its entire genome. Thus, HPV related sequences that have different degrees of

similarity over their length to the most closely related HPV sequence in GenBank are considered as possible chimeras (i.e. it is assumed that they contain parts of different HPVs). The algorithm checks these chimeras by dividing the sequence into three equal segments. If at least one of the segments has less than 90% similarity, another segment has more than 90% similarity and the difference between these segments is more than 5% (e.g. if segment 1 is 88% similar and segment 2 is 94% similar) the sequence is considered as "possibly chimeric". This approach has been extremely valuable when analyzing HPV genotypes, however, it cannot be used for viruses that frequently rearrange parts of genomes with each other (e.g. Anelloviruses) and other algorithms must be developed [130].

Taxonomic classification of metagenomic reads can be performed by similarity and/or non-similarity-based methods. One of the most famous similarity-based taxonomic classifications is performed by NCBI BLAST, where sequences are compared to known genomes. However, a large part of the sequencing reads from *de novo* sequencing projects are classified as unknown [31, 130] due to incompleteness of public sequence databases or drawbacks of NGS technologies such as short read lengths and sequencing errors. Therefore, more sensitive algorithms, such as BLASTx and tBLASTx searches are conducted against protein databases after the BLASTn search on the nucleotide level. An example of this, would be to subject assembled contigs to taxonomic classification by comparing them against GenBank nucleotide database using paracel blast (www.strikingdevelopment.com) BLASTn to classify them as a) previously known sequences, b) related to previously known sequences, or c) unrelated to previously known sequences.

One of the biggest challenges for bioinformatics analysis is the taxonomic classification of NGS data as many of the sequences, especially those belonging to viruses, have no homologs in the public databases or these homologs are highly divergent [136]. In BLAST searches, sequences might have multiple matches and to classify these sequences, several methods have been developed. One of the most frequently used is called MEGA [137]. This method finds the 'Lowest Common Ancestor' node of all matching sequences in the phylogenetic tree, reducing the risk of false positive matches. However, MEGA might produce false negative results by discarding sequences that do not satisfy user-defined cut-offs.

In metagenomic samples, genome size is related to the number of reads and thus, MEGAN is suboptimal for quantitative metagenomic analyses. Nevertheless, other tools have been developed to address this issue. The GAAS (Genome relative Abundance and Average Size) tool [138] iteratively weighs each reference genome for all matching reads and normalizes the number of reads to the length of their genomes. GRAMMy (Genome Relative Abundance estimates based on Mixture Model theory) are another useful tool that models read assignment ambiguities, genome size biases and read distributions along the genomes on a unified probabilistic framework [139]. Both tools estimate similarities from the alignment qualities of the reads to the reference genomes but not from the reference genomes directly. Thus, they are very valuable for divergent genomes, but they are suboptimal in case there are highly similar genomes in the reference databases. The Genome Abundance Similarity Correction (GASiC) considers reference genome similarities to correct the observed abundances estimated via read alignments [140].

1.7 SIGNIFICANCE OF THE STUDY

The development of sequencing technologies that perform high throughput sequencing together with the possibility of performing an unbiased approach, allows for the accurate identification of all microorganisms that are present in large numbers of biological specimens.

The experiments carried out in this thesis have analyzed viral DNA present in skin tumors and the findings have contributed significantly to the knowledge of HPV genotypes. Until 2004, only 106 HPV types had been completely cloned, sequenced and given an official number at the International HPV Reference Center. However, during the last 10 years, 99 new types have been recognized (http://www.hpvcenter.se, accessed on 2016-01-15) and the number of putative novel HPV types is continuously growing. Our group has detected a total of 360 putative novel HPV types (called SE types) so far, all of them belonging to the genera *alpha*, *beta* and *gamma* (Figure 8) and one of them, HPV 197 (SE46) was found to be the most common genotype present in skin tumors.

The resulting “microbiological sequence atlas” obtained in this thesis can be used for large-scale molecular epidemiological analysis on whether any particular infection is regularly present in any specific form of skin cancer. Creating a solid and comprehensive basis for advancing knowledge in this area is important. Furthermore, if extensive sequencing efforts fail to find any infection present in cancer tissue, this would point to other hypotheses (e.g., a role of immunosurveillance in the recognition and elimination of precancerous lesions, regardless of cancer etiology, and in cancer cell gene expression) [15]. If properly carried out, therefore, also negative findings for viral sequences can provide useful clues about the origins of human skin cancer [16] .

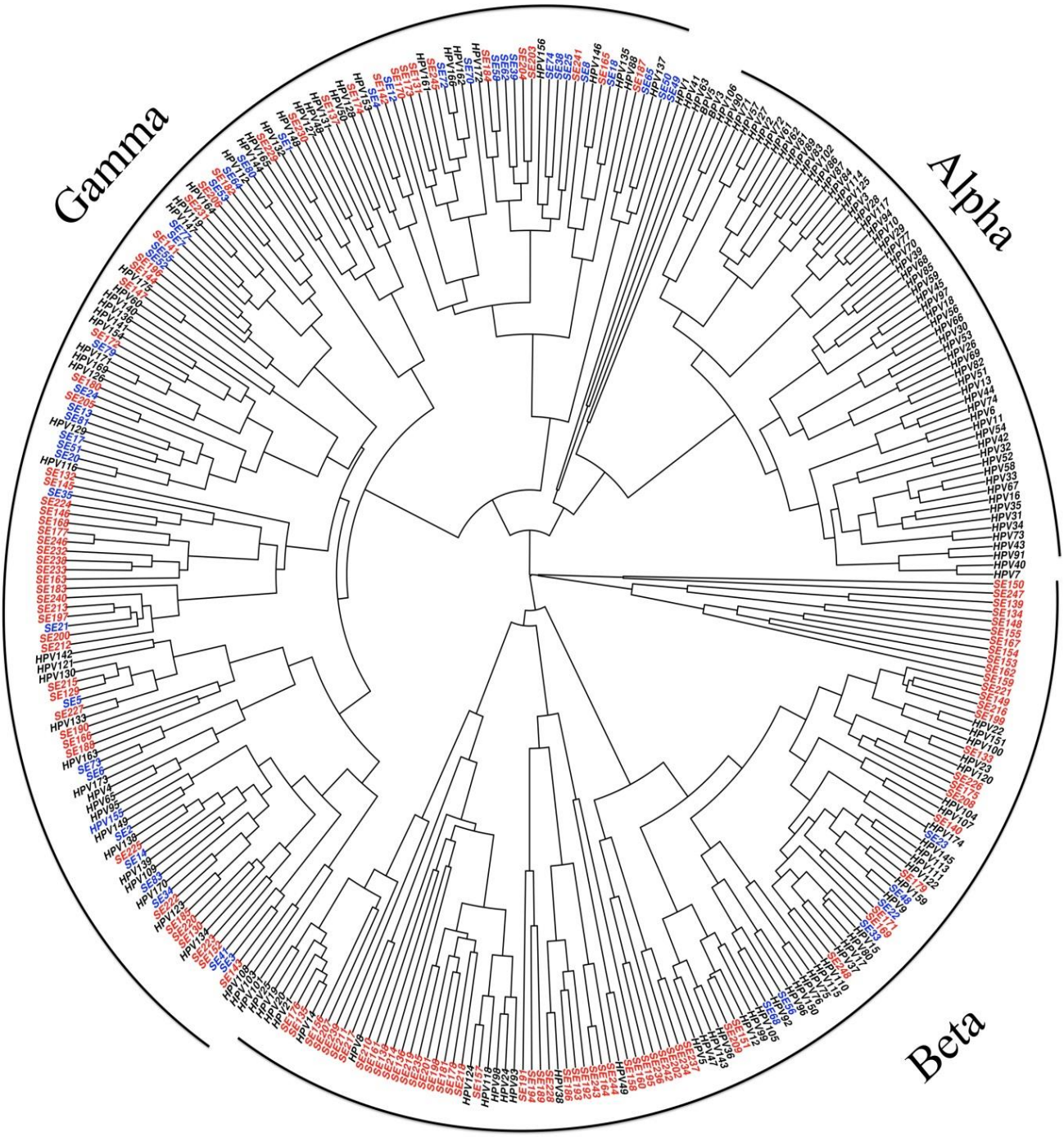

Figure 8: Bayesian phylogenetic tree based on the L1 part of the complete 164 established HPV types (+ bovine papillomavirus type 3 and type 5) and 160 putative novel HPV types (SE-types) that were >400 bp or contained >200 bp of the 3´-end of the amplimer. SE-types discovered using 454GSFLX and using Illumina MiSeq are presented in blue and red colors, respectively.

2 SUMMARY OF PUBLICATIONS

2.1 AIMS

The purpose of the studies included in this thesis, was to:

Paper I: Next generation sequencing for HPV genotyping.

> Validate a high throughput sequencing method (454 GS technology, Roche) for PCR amplimers in order to detect and type HPV and compare it with the HPV genotyping method used by the WHO HPV LabNet global reference laboratory.

Paper II: Diversity of HPV in skin lesions.

> Explore detection of cutaneous HPV genotypes in different skin lesions with optimized chemistry and design of primers.

Paper III: Deep sequencing extends the diversity of HPVs in human skin.

> Investigate the presence of additional viruses in skin lesions using a, compared to previous papers, deeper sequencing technology (Illumina technology).

Paper IV: Human papillomvirus type 197 is commonly present in skin tumors.

> To explore which HPVs are most commonly detected in skin cancers when a sequencing method that is independent of PCR (not biased to any particular sequence) is used.

Paper V: Does Human Papillomavirus-Negative Condylomata Exist?

> Analyze samples from "HPV-negative" condylomata with metagenomic sequencing to investigate which viruses these samples contain (if any).

2.2 MATERIALS AND METHODS

2.2.1 Study material

Paper I: Cervical, urethral and prostate specimens

- Cervical samples:

  Cervical samples (n=62) were selected from Swedescreen, a population-based randomized controlled trial of HPV DNA testing in primary cervical screening [141]. Specimens were collected through endo/ectocervical sampling with a cytobrush. The brush was swirled in 1 mL sterile 0.9% NaCl and samples were immediately frozen.

- Urethral specimens:

  Urethral specimens (n=14) were selected from a university outpatient clinic in St. Petersburg. Samples were collected with a urinary swab that was first inserted into the urethra and rotated 180° right- and leftwards. The swab was rinsed in 1000 µl phosphate buffer and then stored at -20 °C.

- Prostate samples:

  Prostate samples (n=11) were selected from the same university outpatient clinic in St. Petersburg. Specimens were collected after patients were asked to urinate and a digital rectal examination with massage of the prostate was done. The prostate secretion dropping from the urethra was stored at -20 °C.

Frozen cervical, urethral and prostatic specimens were thawed and centrifuged. The pellet was dissolved and used for DNA extraction by a freeze-thaw-boiling procedure [142].

Paper II and III: Skin samples

Two different patient series were used for both papers:

- SCCs (n=119), AKs (n=114), BCCs (n=117), KAs (n=8), seborrheic keratosis (n=46) and two benign lesions (one prurigo nodularis and one benign hyperkeratoic skin lesion) were collected from a hospital-based study in Sweden and Austria. All patients provided four different samples: a swab sample and a

biopsy of the lesion as well as a swab sample and a biopsy from healthy skin. Swabs were collected by a pre-wetted (0.9% NaCl) cotton-tipped swab and biopsies were taken after tape-stripping the skin surface to remove possible environmental contamination. DNA biopsies were extracted using a phenol-free method [49] and with MagnNA Pure LC using the Total Nucleic Acid kit (Roche) whereas the swab samples were extracted by a freeze-thawing [49].

- Fresh frozen KA biopsies (n=92) were included in the study from the Department of Dermatology and Plastic surgery at the Norwegian National Hospital, Oslo, Norway. The DNA was extracted using the QIAmp DNA Minikit (Qiagen) [143].

Paper IV: Skin samples

Two different patient series were used for this paper, formalin-fixed paraffin-embedded blocks and fresh frozen biopsies:

- Formalin-fixed paraffin-embedded (FFPE) SCC biopsies (n=24) were part of a larger study containing 130 SCCs, 378 BCCs and 157 other NMSCs at the Department of Pathology at Malmö University Hospital, Sweden. The paraffin blocks were sectioned, de-paraffinized using xylene and DNA was extracted by a phenol-free method [49].

- SCCs (n=17), AKs (n=22), BCCs (n=3), KAs (n=8) and SCCs in situ (n=17) were part of a Swedish hospital-based study of NMSCs, and premalignant and benign lesions. All patients donated 2 swab samples and 2 biopsies and DNA was extracted as described above.

Paper V: Condyloma specimens

Condyloma swab samples were collected from 703 patients visiting the Centre for Sexual Health in Malmö, Sweden between 2006 and 2009. Swab samples were collected with a pre-wetted (0.9% NaCl) cotton-tipped swab rolled over the condyloma and stored in 1 mL saline. Samples were then pelleted and cell suspension was DNA extracted with MagNA Pure LC using the Total Nucleic Acid Kit (Roche). In a previous study [144], the samples were subjected to MGP amplification followed by Luminex genotyping. Forty swab samples of apparently "HPV-negative" condylomata were included in this Paper.

2.2.2 Methods

2.2.2.1 Sample adequacy

Sample adequacy was assessed by amplification of the beta-globin gene with real-time PCR.

2.2.2.2 HPV Amplification:

HPV specific amplification

In Paper I, two different sets of primers were used for HPV amplification: MGP and PGMY primers. Both general consensus primers amplify a region within the L1 gene (the MGP targeted region lies inside the region amplified by PGMY primers) and were designed to improve amplification of mucosal HPV types.

MGP primers consist of five forward and five reverse primers which have a few modified nucleotides each compared to the general primers GP5+/6+ for better annealing to different HPV types [73]. The resulting HPV amplicon comprises approximately 160 bp. The PGMY primers PCR system is a set of 18 defined primers, which were designed to improve mucosal HPV type amplification of MY09/11 primers. These primers amplify a 450 bp region within the L1 gene.

While HPV genotyping using hybridization probes was carried out after MGP amplification of specimens (as we followed the same method and protocols that are validated and used at the WHO HPV LabNet global reference laboratory in Sweden), the 454 sequencing technology required a longer length of amplimers and thus, the HPV general PGMY primer set was chosen for viral amplification. PGMY primers and amplification protocols are also validated and included in the WHO Human Papillomavirus Laboratory Manual [145].

In Papers II, III and IV, another primer set was used, FAP59/64. These primers were designed to amplify most currently known cutaneous papillomavirus, and to generate an amplicon of about 480 nucleotides in the L1 ORF. The broad HPV amplification is due to the degeneration of primers instead of multiplexing. As FAP primers are normally used for skin samples, most of the FAP amplicons are usually expected to belong to *beta* and *gamma* genera. Some HPV types among *alpha* genus as well as those belonging to *mu* and *nu* genera might escape amplification due to mismatches with the primers (Table 5). Therefore, a new FAP6085/FAP64 primer system was designed in order to amplify and detect those "escaping" HPV types (Paper II).

		FAP59																				
HPV type	Genus	T	A	A	C	W	G	T	I	G	G	I	C	A	Y	C	C	W	T	A	T	T
HPV 16	*Alpha*	-	T	G	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	-
HPV 5	*Beta*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HPV 50	*Gamma*	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HPV 1	*Mu*	-	-	-	A	-	-	-	-	A	-	-	T	G	-	-	-	-	-	T	C	-
HPV 41	*Nu*	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-

		FAP64																							
HPV type	Genus	G	A	T	G	G	I	G	A	I	A	T	G	D	B	W	G	A	T	A	T	W	G		G
HPV 16	*Alpha*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	C	-	-		-
HPV 5	*Beta*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	-		-
HPV 50	*Gamma*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
HPV 1	*Mu*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	-	-	-	-	-	-		-
HPV 41	*Nu*	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	-		-

Table 5: Alignment of the FAP59 and FAP64 primer sequences with the corresponding region in the L1 ORF of various HPV types. Lines and characters represent identical and mismatched nucleotides, respectively. Degenerate nucleotides of primers: W= T, A; I= inosine; Y= C, T; D= A, G, T; B= G, C, T.

Long PCR amplification

In Papers IV and V, complete genome sequences from putative novel types were detected by the sequencing platforms and long PCR amplification was performed to amplify these whole genomes in different overlapping fragments. The PrimeSTAR GXL DNA Polymerase kit from TAKARA was used for this purpose and individual PCR-programs were set for each fragment according to the manufacturer´s instructions. Specific primers were designed for each genome sequence.

Unbiased approach: WGA

In Papers III, IV and V, we attempted to perform an unbiased approach, not dependent on prior sequence information, and detect all DNA organisms present in skin samples by performing a whole genome amplification instead of a specific HPV PCR. DNA was amplified using the IllustraTM Ready-To-Go™ GenomiPhi™ DNA Amplification Kit (GE Health Care, United Kingdom) according to the manufacturer's instructions, with some modifications.

Five microliters of sample were diluted with 20 μL PCR-Grade water and 25 μL of denaturation buffer. The mixture was incubated at 95°C for 3 min and then cooled on ice. For the amplification reaction, 50 μL of the denatured samples were added to the Ready-To-Go GenomiPhi lyophilized cake. Afterwards, samples were incubated at 30°C for 7 hours and inactivated at 65°C for 10 min. Products were dissolved by diluting samples in the ratio 1:1 in PCR-Grade water or TE-buffer.

2.2.2.3 HPV Detection

Luminex system

In Paper I, HPV was detected and genotyped using a multiplex bead-based hybridization method with Luminex, which is the HPV genotyping method used by the WHO HPV LabNet global reference laboratory.

The system consists of type-specific probes bound to fluorescent beads and a 2-color laser which allows recognizing HPV types by detecting the fluorescence of streptavidin-F-Phycoerythrin (identifying that hybridization occurred) and the color of the bead (identifying the HPV type).

Beads with probes for 13 oncogenic (high-risk, HR-HPV) types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68a and 68b) and 23 non-oncogenic types (6, 11, 26, 30, 40, 42, 43, 53, 54, 61, 66, 67, 70, 73, 74, 81, 82, 83, 86, 87, 89, 90 and 91) were used. Furthermore, two "universal" probes were included in the study, aiming to represent other HPV types present in the specimens as well as those HPV types with point mutations that might not bind to their specific probes [146, 147].

For Paper II, the Luminex assay was performed to screen more than 1000 skin samples for 44 putative novel HPV types. Unique probes were designed for each novel sequence and cross-reactivity was investigated. Probes were excluded if cross-reactivity was detected.

Sequencing

GS Junior technology was chosen as the sequencing platform for Papers I and II. However, new sequencing technologies with a deeper throughput came to the market and Papers III, IV, and V use protocols and sequencing data obtained with Illumina technology. A comparison of results from different sequencing technologies when sequencing the same samples can be seen in Paper III.

Real-time PCR

In Paper IV, real-time PCR was used as a screening method to identify a putative novel HPV type (HPV 197) in all specimens included in the study. Specific primers and probes were designed and an HPV 197 plasmid was used as a positive control at different dilutions (from 100,000 to 0.5 copies/μL). Real-time PCR was performed twice for each sample to confirm results and in case of ambiguity, real-time PCR was repeated a third time.

2.2.2.4 New SE types and Cloning HPV types

Nearly 100 sequences representing putatively novel HPV types (SE types) were detected in Papers II, III, IV and V. Usually, complete SE sequences of approximately 440 bp were obtained after sequencing FAP PCR amplimers. Each sequence was blasted against GenBank as well as against the rest of the "novel" HPV sequences found to avoid overestimation of "novel" HPV types. However, shorter sequences within the FAP region were also detected, most probably due to the DNA breakage or trimming in the bioinformatics step which removes low quality sequences. To avoid overestimation of putative new HPV types, phylogenetical analysis was restricted to sequences that were complete or almost complete fragments or contained >200 bp of the 5´-/ 3´-end of the amplimer, in order to eliminate the possibility that non-overlapping sequences might derive from the same virus.

All these putative types with partial sequences were deposited in GenBank, but were not designated an official HPV type number as they were not cloned, submitted to the International HPV Reference Center, and completely sequenced.

When performing WGA, a fragmentation step was performed when preparing the library for sequencing. Consequently, the HPV sequences obtained from one specimen might belong to different regions (E6, E2, L1, etc) and it could not be confirmed if non-overlapping sequences belonged to the same HPV type (as multiple HPVs presence is common). To avoid overestimation of putative new HPV types, all "novel" sequences (SE types) that did not overlap but shared the same closest hit in BLAST, were considered to be the same SE type.

In Papers IV and V, complete genome sequences were obtained for some novel HPV types (HPV 197, 200, 201 and 202) and these types were cloned. Long-PCR was performed as explained above and all fragments were gel-purified and cloned using the Zero Blunt VR TOPO VR PCR Cloning kit (Invitrogen) using the pCRTM-Blunt II-TOPO VR vector.

2.2.2.5 Summary of methods throughout the papers

Paper I: Next generation sequencing for HPV genotyping

In order to validate and optimize a sequencing method for HPV detection, cervical, urethral, and prostate samples that had already been analyzed for HPV presence using Luminex technology (the HPV genotyping method used by the WHO LabNet global reference laboratory) were sequenced with the 454 technology (Roche) and the results were compared.

Sensitivity of the sequencing method was investigated using a pool of plasmids (WHO HPV LabNet global proficiency panel) at different concentrations and reproducibility was assessed by preparing libraries and sequencing a group of specimens twice.

Paper II: Diversity of HPVs in skin lesions

After optimization and validation of the 454 method (Paper I), different pools of skin samples (total 326 specimens) were subjected to sequencing for HPV detection using the same technology, 454 Titanium Chemistry with GS Junior.

A general set of degenerated HPV primers (FAP59/64) was chosen for amplification due to its higher specificity for HPV cutaneous types. Furthermore, a new set of primers (FAP6085/FAP64) was designed, validated with different plasmids, and tested in the same specimens in order to be able to amplify a broader number of cutaneous HPV types (*alpha*, *mu* and *nu* types), especially those where the FAP59/64 had mismatches.

Unique probes were designed for 44 novel subgenomic sequences which had been detected in these skin pools in a previous study [33]. The Luminex assay was performed to investigate the prevalence of the sequences in various skin samples.

Paper III: Deep sequencing extends the diversity of HPV in human skin

The same FAP59/64 amplimer products used in Paper II together with a pool of whole-genome amplified swab samples sequenced with GSFLX (Roche) and Ion Torrent PGM technologies in previous study [63] were sequenced using a deeper throughput technology, Illumina (MiSeq platform).

Libraries for amplimers were prepared using the TruSeq Nano DNA Sample Preparation kit (Illumina), omitting fragmentation, end-repair, and size selection, as specimens consisted of approximately 450 bp long PCR products. Whole-genome amplified material was tagmented and sequenced following the Nextera DNA Sample Preparation kit.

Paper IV: HPV type 197 is commonly present in skin tumors

In order to find all DNA organisms present in skin tumors and their prevalence, extracted DNA from fresh frozen biopsies and FFPE blocks from skin lesions were individually sequenced using both MiSeq and HiSeq instruments (Illumina) after whole genome amplification. A HPV consensus PCR (FAP 59/64) was also performed in order to compare the results of both types of amplification.

A complete genome from a putative novel type (SE46) was detected in various specimens that had been amplified following whole genome amplification. Presence of this novel type was investigated using real-time PCR in all samples. The HPV genome was amplified, cloned and sequenced by Sanger sequencing to confirm the sequence. Clones were sent to the International HPV Reference Center (Karolinska Institutet, Sweden) that confirmed the DNA sequence and assigned the submitted clones the novel type number HPV 197.

Paper V: Does Human Papillomavirus-Negative Condylomata Exist?

Forty samples of MGP-PCR "HPV-negative" condylomas from a previous study [141] were whole genome amplified, tagmented, and individually sequenced using both MiSeq and NextSeq instruments (Illumina technology). Sequencing results from Illumina MiSeq were used to evaluate DNA library quality before sequencing them on Illumina NextSeq500 which provided about 2.7 Gb sequencing depth per sample with a read length of 150 bp paired-end.

The complete genome sequences were obtained for 6 previously unknown HPV sequences, putatively representing new types. Three of them, were amplified and cloned from their respective specimens in two overlapping fragments each and sequenced by Sanger sequencing to confirm the sequences. Clones were sent to the International HPV Reference Center (Karolinska Institutet, Sweden) that confirmed the DNA sequences and assigned the submitted clones the novel type numbers HPV 200, HPV 201 and HPV 202.

2.3 RESULTS AND DISCUSSION

2.3.1 Paper I

NGS was evaluated as an HPV genotyping platform and compared to the HPV genotyping method used by the WHO HPV LabNet global reference laboratory. Sensitivity and reproducibility of the sequencing technology were also studied.

The sensitivity analysis showed that GS Junior presented a higher sensitivity compared to Luminex. A pool of 15 different plasmids was used to perform this experiment and 13/15 were detected at an input of 10 copies. Nevertheless, the numbers of reads detected were not correlated with the actual amount of virus present. This might be explained by the performance of the PCR protocol. All 15 plasmids were amplified as a pool, instead of performing individual PCRs and then pooling the amplicons. Amplifying a pool might translate into competition of HPV plasmids in order to bind to the primers. Those HPV types that bind better to the primer sequences would be more efficiently amplified and consequently may deplete PCR reagents than those HPV types that present mismatches with the primer sequence. The objective of pooling all plasmids at the same time instead of performing individual reactions was to reflect the situation found in real specimens, where multiple infections are common. Junior detected multiple infections with almost no limit in the number of types possible to detect in a sample (15 different types were detected in the sensitivity analysis).

Most of the samples tested (72.7%) for reproducibility of the method showed perfect concordance for all types, even with multiple infections containing up to four types. Partial concordance was found in 6/33 samples, where genotypes with a low number of reads were not detected in one of the runs and three samples presented discordance (one genotype was detected with >50 reads in one run, but not in the other).

GS Junior appeared to be an adequate and efficient tool for HPV genotyping. Sixty percent (36/60) of the samples tested with both genotyping methods showed perfect concordance. GS Junior detected more HPV types in 12 specimens (20%) while Luminex did in 8 (13.3%).

Most genotypes that were not detected with the sequencing technology included HPV 68a, 87, 90 and 91. As described in the Materials and Methods section, different primers were used for each method. When analyzing genome sequences of the genotypes not detected with GS Junior together with the primer sequences, PGMY primers presented many genotype binding mismatches (Table 6). Therefore, these genotypes might not have been efficiently amplified, especially in those cases where other genotypes that bind better to the primers were present in the same specimens. Broader or additional general primers would improve the sequencing method.

Overall, GS Junior presented three main advantages when comparing the results to those obtained from the HPV genotyping reference method (Luminex system). First, the sequence technology was able to detect all HPV types present in a sample while Luminex detected only those that were represented by specific hybridization probes. For example, HPV 32-positivity was detected in one specimen with the sequencing platform while Luminex considered it to be negative due to lack of a specific probe designed for this genotype.

Secondly, mis-typing due to cross-reactivity was non-existent with the sequencing technology. One specimen was genotyped as HPV 114 with GS Junior while it was genotyped as a multiple infection of HPVs 83 and 86 with Luminex. These genotypes are phylogenetically related to HPV 114 and thus, a cross-reaction might have occurred as a specific probe for HPV 114 was not designed.

Lastly, GS Junior enabled the user to obtain the entire sequence of the amplimer that might be useful to look for point mutations or variations for epidemiological studies and surveillance of HPV infection.

HPV 90	G	A	C	C	A	A	T	T	C	C	C	T	C	T	T	G	G	C	A	G
PGMYF	G	A	T	C	A	G	T	T	T	C	C	T	T	T	G	G	G	A	C	G
PGMYG	G	A	T	C	A	G	T	T	T	C	C	T	T	T	A	G	G	T	C	G
PGMYH	G	A	T	C	A	G	T	T	T	C	C	T	T	T	T	G	G	A	C	G
PGMYI	G	A	T	C	A	G	T	T	T	C	C	C	C	T	T	G	G	C		
PGMYJ	G	A	T	C	A	G	T	A	T	C	C	T	T	T	G	G	G	A	C	G
PGMYK	G	A	T	C	A	G	T	A	T	C	C	C	C	T	T	G	G	A	C	G
PGMYL	G	A	T	C	A	A	T	T	C	C	C	T	T	T	A	G	G	T	C	G
PGMYM	G	A	T	C	A	A	T	T	T	C	C	A	C	T	A	G	G	T	C	G
PGMYN	G	A	T	C	A	A	T	A	T	C	C	C	C	T	T	G	G	T	C	G
PGMYP	G	A	T	C	A	G	T	T	T	C	C	G	T	T	G	G	G	C		
PGMYQ	G	A	C	C	A	G	T	T	T	C	C	C	T	T	G	G	G	T	C	G
PGMYR	G	A	C	C	A	G	T	T	T	C	C	T	T	T	A	G	G	A	C	G

Table 6: HPV 90 and the reverse primers binding sequences. Highlighted nucleotides show mismatches with the HPV 90 genomic sequence.

2.3.2 Paper II

Roche sequencing platforms allowed for the possibility of performing bidirectional sequencing and the Titanium chemistry improved quality and read length of sequencing data. Both characteristics were assessed and used for sequencing FAP59/64 amplimers

from 326 skin samples in three different pools – fresh frozen biopsies from SCCs and AKs (A), fresh frozen biopsies from KAs (B) and swab samples from SCCs and AKs (C) - which had already been sequenced without both of the new technology advances [33].

It was speculated that the HPV general primer system FAP might have reached its limit in detecting new viruses [48], but our results showed the high effectiveness of pre-amplication by general primer PCR followed by bidirectional sequencing. Overall, a total of 273 different HPV types or putative types were identified, 47 of those sequences belonging to novel putative types.

The pool with swab samples contained most of the HPV sequences, being more frequently found those HPVs which belonged to genus *gamma* papillomavirus, followed by those in the genera *beta* papillomavirus and *alpha* papillomavirus. Presence of anogenital oncogenic HPV in skin samples has been reported [61, 148] and contamination of the skin by viruses originating from mucosal surfaces, mediated by the fingers has also been confirmed [149, 150]. Fresh frozen biopsies were taken after tape-stripping the skin surface [62], to reduce possible contaminating viruses, and only a single *alpha* papillomavirus (HPV 94) was found, suggesting that its presence might represent an infection. However, swab samples were collected without tape-stripping the skin, and therefore, higher presence of *alpha* papillomavirus was most probably due to contamination of the skin.

HPVs from *mu* and *nu* genera were not detected. As the 3 genotypes (HPV 1, 41, and 63) included in these genera are distantly related to other HPV genera, novel primers were designed to detect those specific HPV types as well as to investigate whether the broad detection of HPV types could be further improved (*alpha*, *beta* and *gamma* genera). This novel combination of primers did not detect *Mu*- or *Nu*- papillomaviruses either. The FAP59/64 primer set detected 76 known HPV types, 117 previously known putative types, and 35 subgenomic sequences representing putatively novel types. Novel primers combination detected 33 known types, 45 previously known putative types, and 12 subgenomic sequences representing putatively novel types. To avoid overestimation of putative new HPV types, phylogenetical analysis was restricted to those sequences that were complete or almost complete fragments or contained >200 bp of the 5-end of the amplimer (to discard the possibility that non-overlapping sequences might derive from the same virus) (Figure 9).

The three pools were re-sequenced using 121 MIDs and altogether sequences from 283 different known and putative HPV types were detected. In pool A, 93 HPV types/putative types were detected. 3 HPV types were detected by all MIDs while up to 20 types/putative types were detected by more than 50% of the analysis. Re-sequencing pool B revealed the presence of 190 known and putative HPV types. Only 6 HPV types were detected by all MIDs and 59 types/putative types were detected by more than half of the re-sequencings. Pool C showed the presence of 174 HPV types (including known and putative types) of which 57 were detected by all MIDs and 90 were detected by more than 50% of the analysis.

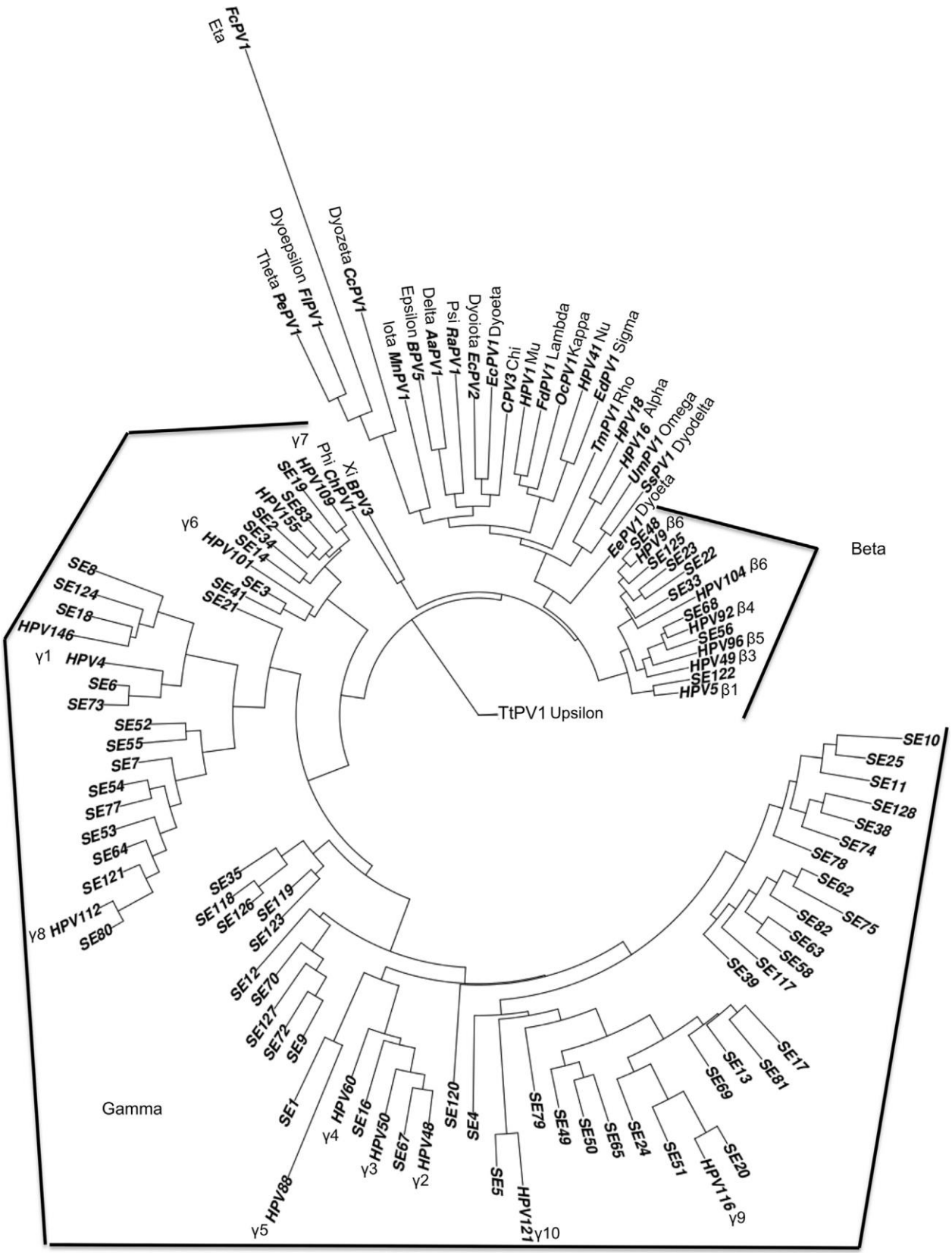

Figure 9: Maximum likelihood tree based on 42 of the novel putative SE sequences identified in this study and L1 from representative known PV types from all other genera. All SE sequences belonged to *gamma* and *beta* genera.

A clear distinction was observed between those HPV types that were detected by all MIDs (and are probably present at high abundance) and those HPV types that were only picked by a few MIDs (and are therefore probably present at much lower copy numbers). High throughput sequencing was able to detect HPV types present in very low copy numbers that would probably otherwise have been missed in the cloning step.

The same pools had been sequenced before in the same platform (without using bidirectional sequencing and Titanium chemistry) and 44 subgenomic sequences representing novel types were identified (SE types). With bidirectional sequencing and Titanium chemistry, most of those sequences were detected with a longer read-length (average of 170 bp longer). Obtaining longer sequences of SE types revealed that some of the previously reported subgenomic SE sequences belonged to the same virus (a total of 37 subgenomic novel sequences instead of 44). Furthermore, most sequences which were considered as preliminary in the previous publication due to insertion/deletion errors leading to stop codons in L1, were detected with a higher quality and without containing premature stop codons in the gene.

2.3.3 Paper III

A deeper sequencing performed with Illumina platform (MiSeq) extended the diversity of HPVs found in human skin. The same pools used in Paper II (HPV consensus PCR of fresh frozen biopsies and top of the lesion swabs) were sequenced with MiSeq platform revealing a total of 159 known HPV types (52 established types and 107 known putative types) and 226 sequences of previously unknown putative novel types, all belonging to the *gamma*, *beta*, and *alpha* genera.

Phylogenetic analysis was performed with 160/226 novel putative HPV types clustered together in the *beta* and *gamma* genera, including those HPV amplimers that contained >400 bp or >200 bp and the 3´-end (to avoid overestimation of novel putative types). Two new branches (probable new HPV species), one for *gamma* and one for *beta* were formed with 17 and 9 novel putative HPV types, respectively. However, confirmation by cloning is needed in order to know if these new branches do indeed represent new HPV species.

Analysis results from either SCC or from KA frozen biopsies showed no specific association between specific HPV types with any particular skin disease. Most HPVs detected were present in the different skin lesions with about the same proportions in all pools and more than half of all HPV sequences detected were novel putative types. Swabbing the surface of the skin enables characterization of the diversity of the viruses present, but it is less informative for determining associations with skin diseases as the viruses that are shed from other places on the body might also be detectable in swabs of skin surfaces on rather distant sites on the body (e.g. *alpha* papillomaviruses).

Continuing improvements of the sequencing technology are likely to continue to reveal an extraordinary and expanding diversity of cutaneous HPVs. In contrast to cervical cancer, where HPV 16 and 18 are detected in most cancer cases, in skin lesions, the majority of HPV types found after FAP amplification were only present in one specimen.

Sequencing whole genome amplified products of swab samples in the MiSeq platform (Illumina) identified about 100,000 reads comprising 21 known HPV types, 2 known putative types, and 3 novel putative HPV types (Table 6). This pool of samples had been sequenced before using two other technologies - GS FLX (Roche) and Personal Genome Machine® (PGM) Ion Torrent – and a 256-fold and 35-fold larger number of viral reads was obtained with Illumina technology, respectively.

All HPV types detected with GS FLX and Ion Torrent platforms (14 types in total) were identified with Illumina technology, except for 3 of them that had appeared with a single read each during previous sequencing runs. Furthermore, a known putative type (SE46) that had been detected with only a partial sequence when using GS FLX and Ion Torrent platforms, was detected with an increased sequence depth of 220 times and its complete genomic sequence when using Illumina technology (Table 7).

Increasing sensitivity and throughput may cause assembly algorithms to construct erroneous "chimeric" sequences by assembling two different sequences from different viruses. Genomic recombination has not been described yet for HPVs. However, multiple HPV infections are common and both naturally occurring genomic recombination and PCR-mediated recombination may mislead phylogenetic analysis. A more strict bioinformatics pipeline, which detects and removes putative chimeras, was developed for this study. Hence, some sequences reported in previous studies did not pass the chimera checking step in the data analysis pipeline and thus, were not included when constructing the phylogenetic tree.

	GSFLX	PGM 300 bp	PGM 400 bp	MiSeq
Total reads	121752	912218	381017	23699142
Total viral reads	380	2750	765	98535
Total HPV reads	378	2744	762	98265
Total established HPV types	5	10	5	21
Total known putative HPV types	2	2	2	2
Total novel putative HPV types	0	0	0	3
SE46 reads	22	132	44	4881

Table 7: Number of reads for the HPV types detected in whole genome amplified skin swab samples (n=142) by different sequencing platforms. GSFLX (Roche), PGM (Ion Torrent), MiSeq (Illumina).

Results from swab samples after PCR or WGA revealed that sequencing of amplimers was unquestionably more sensitive (352 vs 26 different HPVs detected, respectively). However, whole genome amplified HPV 16 used as a plasmid control was detectable when present at about 0.04 copies/cell, indicating a high sensitivity for WGA. Furthermore, at least 11 HPVs that were detected in the metagenomic sequencing, were not detected when sequencing HPV general primer PCR amplimers, implying that these viruses were not effectively amplified by the general primers used.

2.3.4 Paper IV

An unbiased approach (WGA amplification) using NGS was performed in 91 specimens with both MiSeq and HiSeq Illumina platforms. To date, most authors have performed NGS on pooled samples, making it difficult to determine if any specific HPV type is particularly common among different specimens. This was the first study that carried out NGS on individual skin lesions.

MiSeq sequencing identified a total of 73M reads while a higher throughput was obtained with HiSeq technology (1,271M reads). Viral reads represented around 0.03% of the total sequences in both platforms and most of them (>90%) were related to HPV. The non-HPV viruses that were detected with MiSeq and HiSeq belonged to Anelloviridae family (torque teno virus).

Overall, HPV sequences were found in 47/91 specimens, representing four established HPV types (HPV 16, 22, 120 and 124), two previously known putative types (HPV isolate 915 F06002KN1 and SE46), and four previously unknown sequences putatively representing new types (SE361, SE364, SE365 and SE366). All HPV types and putative types detected belonged to the genera *beta* and *gamma* papillomavirus, except for HPV 16 (*alpha* type). Presence of anogenital oncogenic types in skin cancers has been confirmed in previous studies [61, 148]. HPV 16-positive skin cancers are not common and most probably, infection or contamination of the skin by viruses originating from anogenital mucosal surfaces, mediated by the fingers, might be the cause of detection [149, 150].

The complete genome of SE46 isolate was cloned from the FFPE specimen with SCC and clones and complete genome sequence were sent to the International HPV Reference Center at Karolinska Institutet in Sweden, where it was re-sequenced and established as a new HPV type, HPV 197 (GenBank accession number KM085343). HPV 197 comprised 7,278 bp and belonged to the *gamma* genus, but demonstrated only 75% similarity to the most closely related type (HPV 178).

Most reads obtained from both platforms, belonged to the novel type HPV 197 (SE46 isolate) which was found in 36/91 samples -22 samples were positive with HiSeq and 23

specimens with MiSeq- and the majority of the viral reads came from a single FFPE SCC specimen (17,000 reads and 505,000 for MiSeq and HiSeq, respectively).

Real-time PCR confirmed the presence of HPV 197 in 31/36 samples that were HPV 197-positive by sequencing as well as in three other samples (1 BCC and 2 SCCs) from which neither MiSeq or HiSeq had detected this novel HPV type. Full quantification of the number of copies/cell was not performed in the real-time PCR experiments. However, an HPV 197 plasmid at different concentrations was used as a positive control and all samples showed a concentration below 0.5 copies/cell, meaning that the detection of a high number of reads of HPV 197 (such as 17,000 or 505,000 reads with both MiSeq and HiSeq platforms in one FFPE specimen) was due to WGA amplification and not to an initial high amount of virus in that particular specimen. To confirm this conclusion, real-time PCR was performed in amplified products and only the FFPE specimen that contained such a high number of reads showed a concentration above 0,5 copies/cell (>100,000 copies/cell).

In accordance to our study, most authors have found HPV types in skin lesions only at very low viral loads (<1 copy/1,000 cells), but studies have not been consistent regarding which specific types are most commonly present (due to the different primers used) [22, 64, 65]. With an unbiased approach we found that most skin tumor specimens contain HPV, with HPV type 197 being the most commonly detected virus. Nevertheless, further studies are needed to resolve the biological significance of this finding.

For comparison, the same samples were individually amplified with FAP59/64 primers, pooled and sequenced with the MiSeq platform. A total of 24 established HPV types, 13 previously known putative types and 3 novel putative types were detected, reflecting a higher sensitivity when sequencing PCR amplimers. However, only HPV 124 and putative novel type SE361 were detected both when sequencing without prior general primer PCR and when sequencing PCR amplicons.

HPV 197, which was detected in 39.6% of samples when performing an unbiased approach, was not detected when carrying out FAP PCR. Analyzing FAP PCR primer target sequences confirmed several mismatches in HPV 197. PCR was more sensitive than the unbiased method, but it is dependent on a particular sequence. Thus, novel sequences and/or novel viruses might not be detected. More unbiased PCR-independent methods to describe which HPV types are most commonly present in skin lesions are needed. It is essential to first identify which known or unknown HPV types are present in the specimens, as studies looking for RNAs for only some of the viruses present might not be analyzing the major infection present.

2.3.5 Paper V

With NextSeq platform, 700M reads were obtained, of which 363,182 were viral reads (0.05% of total sequences). Deeper sequencing enabled the detection of HPV in almost all condylomata. Ninety-one percent of viral reads were related to HPV sequences and they were present in 37/40 specimens (92.5%). All HPV-positive samples had more than 10 HPV reads and based only on established HPV types, 31/40 samples were positive for multiple HPV types, with a maximum of 8 different established types.

Sequencing of apparently HPV-negative specimens from diseases known to be HPV-associated continues to identify a large number of previously unknown HPV sequences. HPV sequences represented a total of 75 different HPV types or putative types within the genera *alpha*, *beta* and *gamma*, out of which 43 represented novel putative HPV types.

Six of the novel putative HPV types represented complete HPV genomes and three of them were cloned and established as HPV types 200, 201 and 202. They belonged to the genus *gamma* papillomavirus, but shared only 79%, 68% and 82% similarity to the most closely related type.

HPV 6 and HPV 180 were the most frequent genotypes, detected in 30 and 18 samples, respectively. HPV 6 should have been detected by PCR in the previous study, as MGP primers are known to efficiently amplify this genotype [73]. Presence of viral variants with genomic substitutions in the sequences targeted by PCR primers or probes might explain this genotype amplification failure. Deep sequencing solved this problem by detecting HPV types without any prior sequence information.

The rest of the virus related reads (9% of viral reads) belonged to Molluscum contagiosum virus (8% of total viral reads) and less than 1% of all viral reads were related to Baculoviridae, Herpesviridae, Microviridae, Mimiviridae, Parvoviridae, Iridoviridae and unclassified viruses. Interestingly, the Merkel cell polyomavirus, which is commonly present in healthy skin, was not detected.

Molluscum contagiosum virus was detected in 24/40 samples and 22 out of those 24 also contained HPV sequences. This leaves only a single condyloma specimen where we found neither HPV, nor Molluscum.

Presence of molluscum contagiosum virus is known to exist concomitantly with HPV [151]. *Beta* and *gamma* HPVs are frequently detected on healthy skin [32, 33, 135, 152] and therefore, presence of these types might be due to biological contamination from adjacent healthy genital skin. Clinical distinction between condyloma and mollusca is not always straightforward. Conceivably some of the condylomata might have been misdiagnosed and probably were mollusca rather than condylomata.

A previous study sequenced these 40 apparently HPV-negative condyloma specimens in pools of 4 samples with 454 pyrosequencing on a GS Junior instrument (Roche) [135]. HPV sequences were detected in half of the pools, detecting a total of 35 different HPV types (13 established HPV types, 1 known putative HPV sequence as well as 21 novel putative HPV sequences, all belonging to the *gamma* genus).

As expected, more HPV types were detected with the deeper sequencing. It confirmed positivity (with a cutoff at >5 reads per sample) for 11 /13 established HPV types detected in the previous study and 13/21 novel putative HPVs. It also generated longer contigs of HPV sequences and thus, it confirmed that the 9 phylogenetically related but non-overlapping fragments (SE92-100) did belong to the same novel HPV type (HPV 200) that was cloned and sequenced in the present study. Furthermore, another 3 putative novel types that had non-overlapping sequence fragments (SE106, SE107 and SE116) were all found to belong to HPV 201, also cloned and sequenced in the present study.

2.4 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

The hypothesis that NMSC might be associated with an infectious agent derives mainly from the fact that immunosuppressed patients show a 100-fold increase in this type of cancer [18]. The first postulate designed by Robert Koch to establish a causative relationship between a microbe and a disease claimed that, "the microorganism should be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms". Throughout this thesis, deep sequencing technology was applied in samples from skin lesions to find all possible DNA pathogens present (both known and previously unknown). At least 95% of detected viral reads were related to HPV.

We used both PCR methods as well as an unbiased approach for HPV detection. PCR was more sensitive than the unbiased method, but it is dependent on a particular sequence. New degenerated primers were designed in order to amplify a broader number of HPVs, based on known sequences from HPV established types belonging to the 5 genera: *alpha, beta, gamma, mu and nu*. Overall, 56 more HPV types (established genotypes, previously known putative types, and sequences representing novel types) were detected with this novel set. However, viruses that were phylogenetically distant might still have escaped from PCR amplification. The characterized novel HPV 197 genotype (found in more than 1/3 of specimens) for example, was only detected by an unbiased approach. Furthermore, more than 90% of condylomata that had been tested by PCR and were "apparently negative" for HPV, were positive for HPV DNA when the unbiased approach was used.

We found almost 100 putative novel HPV types in total, and characterized 4 novel HPV types (HPV 197, 200, 201 and 202, Table 8). Most of the types were detected in very few patients each, and at a very low viral load (below 0.5 copies/μL). It might be debatable whether such low viral copy numbers are biologically relevant to tumor initiation and

maintenance. The role of these viruses does not seem to correspond to that of high-risk anogenital HPV, which commonly persist in high copy numbers in each tumor cell. HPVs found in skin could be due to contamination or carriage rather than the cause of a transient or persistent infection. The prevalence of HPV DNA in tape-stripped biopsies is far lower than that on the surface [62], further supporting a passenger role. Presumably, multiple positive samples of the same HPV type over a period of time would help to elucidate the role of this virus.

In addition to the DNA presence, further studies on viral activity such as detection of HPV RNA (viral replication) or DNA integration would also facilitate the association of HPV with skin cancer. Contrary to cervical cancer, no signs of HPV integration have been observed in NMSC. Using in situ hybridization or RT-PCR to detect HPV mRNA in SCCs, have also failed to demonstrate activity of HPV so far, detecting viral transcripts sporadically at low levels in occasional tumors [153, 154], with many other tumors testing negative.

Transcriptome sequencing emerged in 2013 and has been developing since, but HPV RNA expression in skin lesions has not been reported yet [155]. The latency period between primary infection and development of cancer is usually of 15 to 40 years [156], and thus, it is possible that an infection involved in carcinogenesis could have disappeared long before the cancer has developed.

The genotypes found in healthy subjects are also mostly from the *beta* and *gamma* genera [32, 33, 135, 152]. This fact does not contradict Koch´s postulate, as it is known that most of the infections linked to human cancers are common in the whole human population, while only a proportion develops into cancer [156]. A synergistic function between persistent viral infections of the skin and ultraviolet exposure has been described [157-159]. A study by Hall et al. showed that the combined effects of *beta* papillomavirus and presence of a susceptible phenotype like fair skin or prolonged sun exposure resulted in a greater risk of SCC than either risk factor alone [160].

Searching for infectious agents in human cancer is not an easy task. The work in this thesis has expanded our knowledge of the wide genomic diversity of HPV on the skin, and finds that more unbiased PCR-independent methods are needed to describe which organisms are most commonly present in skin lesions. Further studies to assess viral infections in cancer and elucidate the mechanistic effects are needed.

HPV	Total	E6	E7	E1	E2	E4	L2	L1	
HPV 197 (SE46)	7278	444	285	1824	1179	477	1599	1548	Number of nucleotides*
	-	1-444	419-703	687-2510	2452-3630	2918-3394	3632-5230	5202-6749	Position of ORF in genome
	-	147	94	607	392	158	532	515	Number of amino acids (aa)
	-	61 (54)	67 (54)	75 (71)	74 (66)	75 (59)	70 (66)	75 (79)	HPV178 Gamma-24 % identity on nt level (aa level)
HPV 200 (SE370)	7137	420	282	1797	1206	X	1509	1545	Number of nucleotides*
	-	1-420	417-698	685-2481	2408-3613	X	3613-5121	5132-6676	Position of ORF in genome
	-	140	94	599	402	X	503	515	Number of amino acids (aa)
	-	78 (72)	81 (84)	79 (80)	82 (79)	X	77 (73)	79 (86)	HPV48 Gamma-2 % identity on nt level (aa level)
HPV 201 (SE371)	7291	423	297	1827	1257	549	1554	1527	Number of nucleotides*
	-	1-423	420-716	700-2526	2468-3724	2934-3482	3726-5279	5289-6815	Position of ORF in genome
	-	141	99	609	419	183	518	509	Number of amino acids (aa)
	-	62(47)	64 (50)	66 (51)	67 (54)	-	65 (55)	68 (68)	HPV163 Gamma-20 % identity on nt level (aa level)
HPV 202 (SE372)	7344	432	300	1818	1179	369	1497	1587	Number of nucleotides*
	-	1-432	434-733	717-2534	2470-3648	3041-3409	3650-5146	5155-6741	Position of ORF in genome
	-	144	100	606	393	123	499	529	Number of amino acids (aa)
	-	87 (89)	86 (90)	87 (90)	88 (86)	90 (86)	77 (79)	82 (92)	HPV140 Gamma-11 % identity on nt level (aa level)

Table 8: The 4 novel established HPV types (HPV 197, HPV 200, HPV 201 and HPV 202) genome organizations and their similarities of the open reading frames (ORFs) to the closest related HPV types. *Including STOP codon.

3 ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to everyone who has been involved in this work and has helped me to grow as a person and as a scientist.

My main supervisor:

Joakim Dillner. Thank you for letting me play in the "first league" of research and most of all, for letting me join the Anethofamily. Thank you for your guidance and inspiration.

My co-supervisors:

Emilie Hultin. Thanks for your advice, support and help as well as for always being there for me, both as a professional and as a friend. You are my role model.

Ola Forslund. Thanks for your honesty and inspiring discussions.

Göran Andersson. Thanks for showing me the pathologist´s point of view, the importance of sampling.

My colleagues and friends:

Carina Eklund. The person I always rely on. She who knows all about almost everything. My non-official supervisor.

Camilla Lagheden. I hope this thesis is just the beginning of our new adventures. With you, there are never "problemas en el cielo". My cloning mate, my senior.

Maria Hortlund. Thanks for making my life happier knowing that people like you exist. My angel.

David Bzhalava. Thank you for all the good moments, your patience (especially when I asked for the data 1M times) and your humor. I will always remember you whenever I am in a seminar (Booooxxxx).

Helena Andersson. Thank you for all the help with the paper work. Without you, it might have been impossible to defend this thesis.

Karin Sundström, Helena Lamin, Sara Nordqvist Kleppe, Nasrin Perskvist, Lars Andersson, Vitaly Smelov, Miriam Elfström, Helena Faust and all the members of Dillner´s group.

To my former colleagues:

Miren Basaras. Mi primera tutora. Gracias por haberme enseñado todo lo que sé. Contigo empecé mi aventura en la investigación. Gracias por tu paciencia, tus consejos, por inspirarme y sobre todo por ser una amiga.

Ramón Cisterna. Gracias por darme la oportunidad de adentrarme en el mundo de la Microbiología.

Elixabete Arrese, Mariangel Lozoya.

To my family:

Mis padres. No tengo páginas suficientes para agradeceros todo lo que habéis hecho por y para mí. Esta tesis es vuestra. Es el fruto de vuestro trabajo, paciencia, constancia, responsabilidad, honestidad, sufrimiento en ocasiones y amor incondicional. Espero con ella que estéis tan orgullosos de vosotros mismos como yo lo estoy de vosotros. Es mi primer paso en firme para acercarme a ser la persona que veo en vosotros. Gracias por ser mis padres. Oqm.

Mi hermano. Joseba, estoy muy orgullosa de tenerte como hermano. Esta tesis también es un poco tuya. Mi vida estaría vacía sin ti. Tqm.

Ibai. Mi bixito**.** Gracias por creer en mí, por quererme todos y cada uno de los días, por soportarme a veces, por entenderme y por acompañarme en la vida. Gracias por haberme elegido para compartir los sueños y sobre todo, gracias por el mejor proyecto de nuestra vida juntos, **Ekhiotz**.

To my friends and very important people that had accompanied me through the way.

Thanks **Hasse**. You know why.

To everyone who has been by my side and has believed in me.
Para todos aquellos que han estado a mi lado y han creído en mí.

4 REFERENCES

1. Katalinic, A., U. Kunze, and T. Schafer, *Epidemiology of cutaneous melanoma and non-melanoma skin cancer in Schleswig-Holstein, Germany: incidence, clinical subtypes, tumour stages and localization (epidemiology of skin cancer).* Br J Dermatol, 2003. **149**(6): p. 1200-6.
2. Lomas, A., J. Leonardi-Bee, and F. Bath-Hextall, *A systematic review of worldwide incidence of nonmelanoma skin cancer.* Br J Dermatol, 2012. **166**(5): p. 1069-80.
3. Eisemann, N., et al., *Non-melanoma skin cancer incidence and impact of skin cancer screening on incidence.* J Invest Dermatol, 2014. **134**(1): p. 43-50.
4. Nikolaou, V. and A.J. Stratigos, *Emerging trends in the epidemiology of melanoma.* Br J Dermatol, 2014. **170**(1): p. 11-9.
5. Glass, A.G. and R.N. Hoover, *The emerging epidemic of melanoma and squamous cell skin cancer.* JAMA, 1989. **262**(15): p. 2097-100.
6. Green, A., *Changing patterns in incidence of non-melanoma skin cancer.* Epithelial Cell Biol, 1992. **1**(1): p. 47-51.
7. 2011, C.I.i.S., *The National Board of Health and Welfare.* Sweden, 2013.
8. Tulinius, H., et al., *Cancer in the Nordic countries, 1981-86. A joint publication of the five Nordic Cancer Registries.* APMIS Suppl, 1992. **31**: p. 1-194.
9. Boukamp, P., *Non-melanoma skin cancer: what drives tumor development and progression?* Carcinogenesis, 2005. **26**(10): p. 1657-67.
10. Grulich, A.E., et al., *Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis.* Lancet, 2007. **370**(9581): p. 59-67.
11. Lindelof, B., et al., *Incidence of skin cancer in 5356 patients following organ transplantation.* Br J Dermatol, 2000. **143**(3): p. 513-9.
12. Berg, D. and C.C. Otley, *Skin cancer in organ transplant recipients: Epidemiology, pathogenesis, and management.* J Am Acad Dermatol, 2002. **47**(1): p. 1-17; quiz 18-20.
13. IARC, *Monographs on the Evaluation of Carcinogenic Risks to Humans*, 2011, IARC: Lyon.
14. Zur Hausen, H., *The search for infectious causes of human cancers: where and why.* Virology, 2009. **392**(1): p. 1-10.
15. Schulz, T.F., *Cancer and viral infections in immunocompromised individuals.* Int J Cancer, 2009. **125**(8): p. 1755-63.
16. Moore, P.S. and Y. Chang, *Why do viruses cause cancer? Highlights of the first century of human tumour virology.* Nat Rev Cancer, 2010. **10**(12): p. 878-89.
17. Vajdic, C.M., et al., *Cutaneous melanoma is related to immune suppression in kidney transplant recipients.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(8): p. 2297-303.
18. Vajdic, C.M. and M.T. van Leeuwen, *Cancer incidence and risk factors after solid organ transplantation.* Int J Cancer, 2009. **125**(8): p. 1747-54.
19. Adami, J., et al., *Cancer risk following organ transplantation: a nationwide cohort study in Sweden.* Br J Cancer, 2003. **89**(7): p. 1221-7.

20. Forslund, O., *Genetic diversity of cutaneous human papillomaviruses.* J Gen Virol, 2007. **88**(Pt 10): p. 2662-9.
21. Ekstrom, J., O. Forslund, and J. Dillner, *Three novel papillomaviruses (HPV109, HPV112 and HPV114) and their presence in cutaneous and mucosal samples.* Virology, 2010. **397**(2): p. 331-6.
22. Vasiljevic, N., et al., *Four novel human betapapillomaviruses of species 2 preferentially found in actinic keratosis.* J Gen Virol, 2008. **89**(Pt 10): p. 2467-74.
23. Kullander, J., O. Forslund, and J. Dillner, *Staphylococcus aureus and squamous cell carcinoma of the skin.* Cancer Epidemiol Biomarkers Prev, 2009. **18**(2): p. 472-8.
24. Foulongne, V., et al., *Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing.* PLoS One, 2012. **7**(6): p. e38499.
25. Shope, R.E. and E.W. Hurst, *Infectious Papillomatosis of Rabbits : With a Note on the Histopathology.* J Exp Med, 1933. **58**(5): p. 607-24.
26. de Villiers, E.M., et al., *Classification of papillomaviruses.* Virology, 2004. **324**(1): p. 17-27.
27. Bernard, H.U., et al., *Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments.* Virology, 2010. **401**(1): p. 70-9.
28. Chen, Z., L.B. de Freitas, and R.D. Burk, *Evolution and classification of oncogenic human papillomavirus types and variants associated with cervical cancer.* Methods Mol Biol, 2015. **1249**: p. 3-26.
29. Bzhalava, D., C. Eklund, and J. Dillner, *International standardization and classification of human papillomavirus types.* Virology, 2015. **476**: p. 341-4.
30. Arroyo Muhr, L.S., et al., *Human papillomavirus type 197 is commonly present in skin tumors.* Int J Cancer, 2015. **136**(11): p. 2546-55.
31. Bzhalava, D., et al., *Unbiased approach for virus detection in skin lesions.* PLoS One, 2013. **8**(6): p. e65953.
32. Bzhalava, D., et al., *Deep sequencing extends the diversity of human papillomaviruses in human skin.* Sci Rep, 2014. **4**: p. 5807.
33. Ekstrom, J., et al., *High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions.* Int J Cancer, 2011. **129**(11): p. 2643-50.
34. Boulet, G., et al., *Human papillomavirus: E6 and E7 oncogenes.* Int J Biochem Cell Biol, 2007. **39**(11): p. 2006-11.
35. Munger, K., *The role of human papillomaviruses in human cancers.* Front Biosci, 2002. **7**: p. d641-9.
36. Yim, E.K. and J.S. Park, *The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis.* Cancer Res Treat, 2005. **37**(6): p. 319-24.
37. Narisawa-Saito, M. and T. Kiyono, *Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: roles of E6 and E7 proteins.* Cancer Sci, 2007. **98**(10): p. 1505-11.
38. Hughes M, G.L., *Skin cancer viruses: bench to bedside – HPV, HHV8 and Merkel cell carcinoma virus.* Drug Discovery Today: Disease Mechanisms 2013. **10**: p. 91-94.
39. Boxman, I.L., et al., *Transduction of the E6 and E7 genes of epidermodysplasia-verruciformis-associated human papillomaviruses alters human keratinocyte growth and differentiation in organotypic cultures.* J Invest Dermatol, 2001. **117**(6): p. 1397-404.

40. Giampieri, S. and A. Storey, *Repair of UV-induced thymine dimers is compromised in cells expressing the E6 protein from human papillomaviruses types 5 and 18.* Br J Cancer, 2004. **90**(11): p. 2203-9.
41. Billecke, C.A., et al., *Lack of functional pRb results in attenuated recovery of mRNA synthesis and increased apoptosis following UV radiation in human breast cancer cells.* Oncogene, 2002. **21**(29): p. 4481-9.
42. Jablonska, S., J. Dabrowski, and K. Jakubowicz, *Epidermodysplasia verruciformis as a model in studies on the role of papovaviruses in oncogenesis.* Cancer Res, 1972. **32**(3): p. 583-9.
43. Jablonska, S. and S. Majewski, *Epidermodysplasia verruciformis: immunological and clinical aspects.* Curr Top Microbiol Immunol, 1994. **186**: p. 157-75.
44. Pfister, H., *Chapter 8: Human papillomavirus and skin cancer.* J Natl Cancer Inst Monogr, 2003(31): p. 52-6.
45. Fuchs PG, P.H., *Papillomaviruses in epidermodysplasia verruciformis.* Papillomavirus Rep, 1990. **1**: p. 1-4.
46. Orth, G., et al., *Characterization of two types of human papillomaviruses in lesions of epidermodysplasia verruciformis.* Proc Natl Acad Sci U S A, 1978. **75**(3): p. 1537-41.
47. Orth, G., *Epidermodysplasia verruciformis: a model for understanding the oncogenicity of human papillomaviruses.* Ciba Found Symp, 1986. **120**: p. 157-74.
48. 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, 2010. **397**(1): p. 205-16.
49. Forslund, O., et al., *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, 1999. **80 (Pt 9)**: p. 2437-43.
50. Bouwes Bavinck, J.N., E.I. Plasmeijer, and M.C. Feltkamp, *Beta-papillomavirus infection and skin cancer.* J Invest Dermatol, 2008. **128**(6): p. 1355-8.
51. Antonsson, A., et al., *General acquisition of human papillomavirus infections of skin occurs in early infancy.* J Clin Microbiol, 2003. **41**(6): p. 2509-14.
52. Gottschling, M., et al., *Cutaneotropic human beta-/gamma-papillomaviruses are rarely shared between family members.* J Invest Dermatol, 2009. **129**(10): p. 2427-34.
53. Weissenborn, S.J., et al., *Intrafamilial transmission and family-specific spectra of cutaneous betapapillomaviruses.* J Virol, 2009. **83**(2): p. 811-6.
54. Hsu, J.Y., et al., *Shared and persistent asymptomatic cutaneous human papillomavirus infections in healthy skin.* J Med Virol, 2009. **81**(8): p. 1444-9.
55. Antonsson, A., et al., *Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents.* J Gen Virol, 2003. **84**(Pt 7): p. 1881-6.
56. Antonsson, A., et al., *The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses.* J Virol, 2000. **74**(24): p. 11636-41.
57. Boxman, I.L., et al., *Detection of human papillomavirus DNA in plucked hairs from renal transplant recipients and healthy volunteers.* J Invest Dermatol, 1997. **108**(5): p. 712-5.
58. de Koning, M.N., et al., *Betapapillomaviruses frequently persist in the skin of healthy individuals.* J Gen Virol, 2007. **88**(Pt 5): p. 1489-95.

59. Plasmeijer, E.I., et al., *Persistence of betapapillomavirus infections as a risk factor for actinic keratoses, precursor to cutaneous squamous cell carcinoma.* Cancer Res, 2009. **69**(23): p. 8926-31.
60. Pfister, H. and J. Ter Schegget, *Role of HPV in cutaneous premalignant and malignant tumors.* Clin Dermatol, 1997. **15**(3): p. 335-47.
61. Asgari, M.M., et al., *Detection of human papillomavirus DNA in cutaneous squamous cell carcinoma among immunocompetent individuals.* J Invest Dermatol, 2008. **128**(6): p. 1409-17.
62. 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, 2004. **123**(2): p. 388-94.
63. Mackintosh, L.J., et al., *Presence of beta human papillomaviruses in nonmelanoma skin cancer from organ transplant recipients and immunocompetent patients in the West of Scotland.* Br J Dermatol, 2009. **161**(1): p. 56-62.
64. Weissenborn, S.J., et al., *Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers.* J Invest Dermatol, 2005. **125**(1): p. 93-7.
65. Vasiljevic, N., et al., *Characterization of two novel cutaneous human papillomaviruses, HPV93 and HPV96.* J Gen Virol, 2007. **88**(Pt 5): p. 1479-83.
66. Hazard, K., et al., *Subtype HPV38b[FA125] demonstrates heterogeneity of human papillomavirus type 38.* Int J Cancer, 2006. **119**(5): p. 1073-7.
67. Allander, T., et al., *A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species.* Proc Natl Acad Sci U S A, 2001. **98**(20): p. 11609-14.
68. Duhaime, M.B. and M.B. Sullivan, *Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline.* Virology, 2012. **434**(2): p. 181-6.
69. Duncavage, E.J., et al., *Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue.* J Mol Diagn, 2011. **13**(3): p. 325-33.
70. Depledge, D.P., et al., *Specific capture and whole-genome sequencing of viruses from clinical samples.* PLoS One, 2011. **6**(11): p. e27805.
71. Koehler, J.W., et al., *Development and evaluation of a panel of filovirus sequence capture probes for pathogen detection by next-generation sequencing.* PLoS One, 2014. **9**(9): p. e107007.
72. Wylie, T.N., et al., *Enhanced virome sequencing using targeted sequence capture.* Genome Res, 2015. **25**(12): p. 1910-20.
73. Soderlund-Strand, A., J. Carlson, and J. Dillner, *Modified general primer PCR system for sensitive detection of multiple types of oncogenic human papillomavirus.* J Clin Microbiol, 2009. **47**(3): p. 541-6.
74. Cai, Y.P., et al., *Comparison of human papillomavirus detection and genotyping with four different prime sets by PCR-sequencing.* Biomed Environ Sci, 2013. **26**(1): p. 40-7.
75. Telenius, H., et al., *Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer.* Genomics, 1992. **13**(3): p. 718-25.
76. Zhang, L., et al., *Whole genome amplification from a single cell: implications for genetic analysis.* Proc Natl Acad Sci U S A, 1992. **89**(13): p. 5847-51.

77. Paunio, T., I. Reima, and A.C. Syvanen, *Preimplantation diagnosis by whole-genome amplification, PCR amplification, and solid-phase minisequencing of blastomere DNA.* Clin Chem, 1996. **42**(9): p. 1382-90.
78. Fire, A. and S.Q. Xu, *Rolling replication of short DNA circles.* Proc Natl Acad Sci U S A, 1995. **92**(10): p. 4641-5.
79. Garmendia, C., et al., *The bacteriophage phi 29 DNA polymerase, a proofreading enzyme.* J Biol Chem, 1992. **267**(4): p. 2594-9.
80. Blanco, L., et al., *Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication.* J Biol Chem, 1989. **264**(15): p. 8935-40.
81. Dunning, A.M., P. Talmud, and S.E. Humphries, *Errors in the polymerase chain reaction.* Nucleic Acids Res, 1988. **16**(21): p. 10393.
82. Saiki, R.K., et al., *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.* Science, 1988. **239**(4839): p. 487-91.
83. Dean, F.B., et al., *Comprehensive human genome amplification using multiple displacement amplification.* Proc Natl Acad Sci U S A, 2002. **99**(8): p. 5261-6.
84. Hosono, S., et al., *Unbiased whole-genome amplification directly from clinical samples.* Genome Res, 2003. **13**(5): p. 954-64.
85. Binga, E.K., R.S. Lasken, and J.D. Neufeld, *Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology.* ISME J, 2008. **2**(3): p. 233-41.
86. Polson, S.W., S.W. Wilhelm, and K.E. Wommack, *Unraveling the viral tapestry (from inside the capsid out).* ISME J, 2011. **5**(2): p. 165-8.
87. Lasken, R.S. and T.B. Stockwell, *Mechanism of chimera formation during the Multiple Displacement Amplification reaction.* BMC Biotechnol, 2007. **7**: p. 19.
88. Yilmaz, S., M. Allgaier, and P. Hugenholtz, *Multiple displacement amplification compromises quantitative analysis of metagenomes.* Nat Methods, 2010. **7**(12): p. 943-4.
89. Dichosa, A.E., et al., *Artificial polyploidy improves bacterial single cell genome recovery.* PLoS One, 2012. **7**(5): p. e37387.
90. Wang, J., et al., *Microarray-based evaluation of whole-community genome DNA amplification methods.* Appl Environ Microbiol, 2011. **77**(12): p. 4241-5.
91. Abulencia, C.B., et al., *Environmental whole-genome amplification to access microbial populations in contaminated sediments.* Appl Environ Microbiol, 2006. **72**(5): p. 3291-301.
92. Dinsdale, E.A., et al., *Functional metagenomic profiling of nine biomes.* Nature, 2008. **452**(7187): p. 629-32.
93. Dinsdale, E.A., et al., *Microbial ecology of four coral atolls in the Northern Line Islands.* PLoS One, 2008. **3**(2): p. e1584.
94. Cassman, N., et al., *Oxygen minimum zones harbour novel viral communities with low diversity.* Environ Microbiol, 2012. **14**(11): p. 3043-65.
95. Hewson, I., et al., *Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes.* Appl Environ Microbiol, 2012. **78**(18): p. 6583-91.
96. Willner, D., et al., *Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals.* PLoS One, 2009. **4**(10): p. e7370.
97. Marine, R., et al., *Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome.* Microbiome, 2014. **2**(1): p. 3.

98. Kim, K.H., et al., *Amplification of uncultured single-stranded DNA viruses from rice paddy soil.* Appl Environ Microbiol, 2008. **74**(19): p. 5975-85.
99. Dean, F.B., et al., *Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification.* Genome Res, 2001. **11**(6): p. 1095-9.
100. Lage, J.M., et al., *Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH.* Genome Res, 2003. **13**(2): p. 294-307.
101. Michael, K.M., et al., *Bead-based multiplex genotyping of 58 cutaneous human papillomavirus types.* J Clin Microbiol, 2011. **49**(10): p. 3560-7.
102. Schmitt, M., et al., *Evaluation of a novel multiplex human papillomavirus (HPV) genotyping assay for HPV types in skin warts.* J Clin Microbiol, 2011. **49**(9): p. 3262-7.
103. 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, 1995. **33**(3): p. 690-5.
104. Abreu, A.L., et al., *A review of methods for detect human Papillomavirus infection.* Virol J, 2012. **9**: p. 262.
105. Villa LL, D.L., *Methods for detection of HPV infection and its clinical utility.* Int J Gyn Obst, 2006. **96**: p. 71-80.
106. Eklund, C., et al., *Global improvement in genotyping of human papillomavirus DNA: the 2011 HPV LabNet International Proficiency Study.* J Clin Microbiol, 2014. **52**(2): p. 449-59.
107. Eklund, C., et al., *The 2010 global proficiency study of human papillomavirus genotyping in vaccinology.* J Clin Microbiol, 2012. **50**(7): p. 2289-98.
108. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
109. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.
110. Cox-Foster, D.L., et al., *A metagenomic survey of microbes in honey bee colony collapse disorder.* Science, 2007. **318**(5848): p. 283-7.
111. Korbel, J.O., et al., *Paired-end mapping reveals extensive structural variation in the human genome.* Science, 2007. **318**(5849): p. 420-6.
112. Palacios, G., et al., *A new arenavirus in a cluster of fatal transplant-associated diseases.* N Engl J Med, 2008. **358**(10): p. 991-8.
113. Andries, K., et al., *A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis.* Science, 2005. **307**(5707): p. 223-7.
114. Briggs, A.W., et al., *Patterns of damage in genomic DNA sequences from a Neandertal.* Proc Natl Acad Sci U S A, 2007. **104**(37): p. 14616-21.
115. Green, R.E., et al., *Analysis of one million base pairs of Neanderthal DNA.* Nature, 2006. **444**(7117): p. 330-6.
116. Noonan, J.P., et al., *Sequencing and analysis of Neanderthal genomic DNA.* Science, 2006. **314**(5802): p. 1113-8.
117. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
118. Campbell, P.J., et al., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nat Genet, 2008. **40**(6): p. 722-9.

119. Hillier, L.W., et al., *Whole-genome sequencing and variant discovery in C. elegans.* Nat Methods, 2008. **5**(2): p. 183-8.
120. Salzberg, S.L., et al., *Gene-boosted assembly of a novel bacterial genome from very short reads.* PLoS Comput Biol, 2008. **4**(9): p. e1000186.
121. Srivatsan, A., et al., *High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies.* PLoS Genet, 2008. **4**(8): p. e1000139.
122. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
123. Cokus, S.J., et al., *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008. **452**(7184): p. 215-9.
124. Lister, R., et al., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell, 2008. **133**(3): p. 523-36.
125. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-70.
126. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.
127. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
128. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
129. Bokulich, N.A., et al., *Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.* Nat Methods, 2013. **10**(1): p. 57-9.
130. Bzhalava, D., et al., *Phylogenetically diverse TT virus viremia among pregnant women.* Virology, 2012. **432**(2): p. 427-34.
131. Hutchison, C.A., 3rd, et al., *Cell-free cloning using phi29 DNA polymerase.* Proc Natl Acad Sci U S A, 2005. **102**(48): p. 17332-6.
132. Niu, B., et al., *Artificial and natural duplicates in pyrosequencing reads of metagenomic data.* BMC Bioinformatics, 2010. **11**: p. 187.
133. Gomez-Alvarez, V., T.K. Teal, and T.M. Schmidt, *Systematic artifacts in metagenomes from complex microbial communities.* ISME J, 2009. **3**(11): p. 1314-7.
134. Meiring, T.L., et al., *Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits.* Virol J, 2012. **9**: p. 164.
135. Johansson, H., et al., *Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types.* Virology, 2013. **440**(1): p. 1-7.
136. Fancello, L., D. Raoult, and C. Desnues, *Computational tools for viral metagenomics and their application in clinical research.* Virology, 2012. **434**(2): p. 162-74.
137. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Res, 2007. **17**(3): p. 377-86.
138. Angly, F.E., et al., *The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes.* PLoS Comput Biol, 2009. **5**(12): p. e1000593.
139. Xia, L.C., et al., *Accurate genome relative abundance estimation based on shotgun metagenomic reads.* PLoS One, 2011. **6**(12): p. e27992.
140. Lindner, M.S. and B.Y. Renard, *Metagenomic abundance estimation and diagnostic testing on species level.* Nucleic Acids Res, 2013. **41**(1): p. e10.
141. Naucler, P., et al., *Human papillomavirus and Papanicolaou tests to screen for cervical cancer.* N Engl J Med, 2007. **357**(16): p. 1589-97.

142. Forslund, O., et al., *Population-based type-specific prevalence of high-risk human papillomavirus infection in middle-aged Swedish women.* J Med Virol, 2002. **66**(4): p. 535-41.
143. Forslund, O., et al., *Identification of human papillomavirus in keratoacanthomas.* J Cutan Pathol, 2003. **30**(7): p. 423-9.
144. Sturegard, E., et al., *Human papillomavirus typing in reporting of condyloma.* Sex Transm Dis, 2013. **40**(2): p. 123-9.
145. WHO, *Human papillomavirus laboratory manual. First ed*, ed. A.a. www.who.int/vaccines-documents/. 2009.
146. Schmitt, M., et al., *Homogeneous amplification of genital human alpha papillomaviruses by PCR using novel broad-spectrum GP5+ and GP6+ primers.* J Clin Microbiol, 2008. **46**(3): p. 1050-9.
147. Schmitt, M., et al., *Bead-based multiplex genotyping of human papillomaviruses.* J Clin Microbiol, 2006. **44**(2): p. 504-12.
148. 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, 2003. **63**(21): p. 7515-9.
149. Alam, M., J.B. Caldwell, and Y.D. Eliezri, *Human papillomavirus-associated digital squamous cell carcinoma: literature review and report of 21 new cases.* J Am Acad Dermatol, 2003. **48**(3): p. 385-93.
150. Forslund, O., P. Nordin, and B.G. Hansson, *Mucosal human papillomavirus types in squamous cell carcinomas of the uterine cervix and subsequently on fingers.* Br J Dermatol, 2000. **142**(6): p. 1148-53.
151. Castronovo, C., et al., *Viral infections of the pubis.* Int J STD AIDS, 2012. **23**(1): p. 48-50.
152. Arroyo, L.S., et al., *Next generation sequencing for human papillomavirus genotyping.* J Clin Virol, 2013. **58**(2): p. 437-42.
153. Dang, C., et al., *E6/E7 expression of human papillomavirus types in cutaneous squamous cell dysplasia and carcinoma in immunosuppressed organ transplant recipients.* Br J Dermatol, 2006. **155**(1): p. 129-36.
154. Purdie, K.J., et al., *Human papillomavirus gene expression in cutaneous squamous cell carcinomas from immunosuppressed and immunocompetent individuals.* J Invest Dermatol, 2005. **125**(1): p. 98-107.
155. Arron, S.T., et al., *Transcriptome sequencing demonstrates that human papillomavirus is not active in cutaneous squamous cell carcinoma.* J Invest Dermatol, 2011. **131**(8): p. 1745-53.
156. zur Hausen, H., *The search for infectious causes of human cancers: where and why (Nobel lecture).* Angew Chem Int Ed Engl, 2009. **48**(32): p. 5798-808.
157. Nindl, I., M. Gottschling, and E. Stockfleth, *Human papillomaviruses and non-melanoma skin cancer: basic virology and clinical manifestations.* Dis Markers, 2007. **23**(4): p. 247-59.
158. Jackson, S. and A. Storey, *E6 proteins from diverse cutaneous HPV types inhibit apoptosis in response to UV damage.* Oncogene, 2000. **19**(4): p. 592-8.
159. Jackson, S., et al., *Role of Bak in UV-induced apoptosis in skin cancer and abrogation by HPV E6 proteins.* Genes Dev, 2000. **14**(23): p. 3065-73.
160. Hall, L., et al., *Re: Human papillomavirus infection and incidence of squamous cell and basal cell carcinomas of the skin.* J Natl Cancer Inst, 2006. **98**(19): p. 1425-6.