

From DEPARTMENT OF LABORATORY MEDICINE
Karolinska Institutet, Stockholm, Sweden

STUDIES ON TUMOR VIRUS EPIDEMIOLOGY

Davit Bzhalava



**Karolinska
Institutet**

Stockholm 2014

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden
© Davit Bzhalava, 2014
ISBN 978-91-7549-493-7

To my family

ABSTRACT

The causal relationship between several virus infections and human cancers are well established. However, it is also possible that additional cancers may be caused by known or yet unknown viruses. The present thesis has sought to both further elucidate known relationships between virus and cancer as well as to provide a basis for further exploration in the area of infections and cancer.

Infections during pregnancy have been suspected to be involved in the etiology of childhood leukemias. However, no specific infectious agent is yet linked to the etiology of these diseases. As a basis for further studies in this area, we applied high-throughput next generation sequencing (NGS) technology to describe the viruses most readily detectable in serum samples of mothers to leukemic children. The most common viruses found were TT viruses, including several previously not described TT viruses.

Merkel cell polyomavirus (MCV) is found in Merkel cell carcinoma (MCC), a rare and aggressive neuroendocrine tumor of the skin. To explore whether MCV infection might be associated with additional cancers, we investigated whether MCC patients are at excess risk of other cancers, using population-based Nordic cancer registries. Bidirectional evaluation of excess risk of other diseases among MCC patients revealed that they are at increased risk of other skin cancers as a second cancer, compared to the general Nordic population. Shared causative factors, such as exposure to ultraviolet light and/or MCV infection are among the possible explanations. Also, impact of increased surveillance of the skin should be noted as an explanation of the excess risk.

Cutaneous human papillomaviruses (HPV) are suspected to be involved in the etiology of non-melanoma skin cancer (NMSC). To evaluate whether there are any consistent association between cutaneous HPV infections and skin cancer, we conducted a systematic review and meta-analysis of studies that investigated HPV prevalences among cases of skin lesions and their healthy controls. We found that HPV species Beta-1, Beta-2, Beta-3 and Gamma-1 were more frequently detected in squamous cell carcinoma (SCC) compared to healthy controls.

To provide clues about possible carcinogenicity of 47 mucosal HPV types, out of which 12 are established as causes of cervical cancer, we also investigated the prevalence of 47 mucosal HPV types across the entire range of cervical diagnoses from normal to cervical cancer.

To investigate diversity of HPVs in skin lesions with increased sensitivity, different sample types from different skin lesion were subjected to high-throughput NGS after PCR amplification. Conventional molecular detection methods such as PCR are biased towards the primers used. Thus they might miss viruses that are divergent from the primer sequences. We also investigated whether NGS technology can be used to assess presence of virus DNA in an unbiased manner, both in skin lesions as well as in condylomas that were classified as “HPV negative” by conventional PCR methods.

Unbiased sequencing identified two putatively new HPV types that were missed by NGS after PCR amplification. The advantage of unbiased sequencing over conventional molecular detection methods was further demonstrated in the study of “HPV negative” condylomas. We found several known as well as several putatively novel HPV types in condylomas that were previously found to be HPV negative by PCR.

In conclusion, we have used registry linkage studies, systematic reviews and meta-analyses and modern NGS technology applied to biobanked specimens to extend our knowledge of the epidemiology of cancer-associated viruses and to provide a basis for further exploration in this area.

SAMMANFATTNING

Det är väl känt att flera olika virusinfektioner kan orsaka eller medverka till utveckling av vissa former av cancer hos människan. Det är möjligt att det finns ytterligare cancrar som orsakas av kända eller ännu okända virus. Denna avhandling försöker klargöra kända samband mellan virusinfektion och cancer samt ge en grund för fortsatt forskning inom området infektioner och cancer.

Infektioner under graviditeten har misstänkts vara inblandade i uppkomsten av leukemi hos barn. Hittills har ingen specifik infektion kunnat kopplas till uppkomsten av dessa sjukdommar. Som en grund för fortsatta studier inom detta område använde vi ”Nästa Generation högeffektiv Sekvensering” (NGS) för att detektera de virus som var mest förekommande i serum från mammor till barn som utvecklat leukemi. De virus som detekterades mest var TT-virus, även flera TT-virus som tidigare inte beskrivits.

Merkelcell polyomavirus (MCV) finns i Merkelcellcancer (MCC), en ovanlig och aggressiv neuroendokrin hudcancer. För att se om en MCV infektion kan vara associerad med andra cancrar undersökte vi om MCC patienter hade en ökad risk för andra cancrar med hjälp av populationsbaserade nordiska cancerregister. Tvåvägs utvärdering av ökad risk för andra sjukdomar hos MCC patienter visade att de har en ökad risk för andra hudcancerar som sekundär cancer jämfört med den nordiska befolkningen. Gemensamma orsaksfaktorer, som exponering för ultraviolett ljus och/eller MCV infektion är möjliga förklaringar. Även mer ingående undersökningar av huden kan vara en möjlig orsak till den ökade risk som observerats.

Hudrelaterat Humant Papillomavirus (HPV) tros vara en av orsakerna till uppkomsten av icke-melanom hudcancer (NMSC). För att undersöka om det finns ett konsekvent samband mellan HPV infektioner på huden och hudcancer gjorde vi en systematisk genomgång och metaanalys av studier som undersökt HPV förekomsten i hud lesioner samt i normala kontroller. Vi fann att HPV species Beta-1, Beta-2, Beta-3 och Gamma-1 kunde detekteras mer frekvent i skivepitelcancer (SCC) än hos friska kontroller.

För att ta fram mer information om den möjliga cancerogeniteten hos 47 genitala HPV typer, av vilka 12 är etablerade som orsak till cervixcancer undersökte vi förekomsten av dessa 47 genitala HPV typer inom alla diagnoser för cervix från normal till cervixcancer.

Vi undersökte mångfalden av HPV i lesioner från huden med en metod som ger ökad känslighet jämfört med tidigare studier. Olika provmaterial från olika hud lesioner amplifierades med HPV specifik PCR varefter nästa generation högeffektiv sekvensering gjordes på materialet. Resultaten av konventionella detektionsmetoder som PCR påverkas av de primers som används. Det gör att en metod kan missa virus som har en avvikande sekvens i primer området. Vi undersökte därför om NGS kan användas för att detektera virus DNA utan att först göra PCR både i hud lesioner och i kondylom som tidigare varit HPV negativa med vanliga PCR metoder.

Sekvensering utan föregående PCR identifierade två troligt nya HPV typer som inte hittades med NGS efter PCR amplifiering. Fördelen med denna sekvenseringsstrategi jämfört med konventionella detektionsmetoder visades även i studien av HPV negativa kondylom där flera kända så väl som tidigare okända HPV typer detekterades med NGS men inte med HPV specifik PCR.

Sammanfattningsvis har vi använt studier baserade på registerlänkningar, systematisk genomgång av artiklar och metaanalyser samt modern högeffektiv sekvenserings teknologi på biobanksprover för att utöka vår kunskap inom epidemiologin av cancerassocierade virus och för att ge en grund för fortsatt forskning inom detta område.

აბსტრაქტი

თანამედროვე მეცნიერებისთვის კარგად ცნობილია რომ ადამიანის რამოდენიმე სიმსივნე გამომწვეურია ვირუსური ინფექციებით. თუმცა არსებობს მოსაზრება რომ გაცილებით მეტი სიმსივნეა დაკავშირებული ცნობილ თუ ჯერ კიდევ უცნობ ვირუსურ ინფექციებთან. წარმოდგენილი თეზისი ცდილობს კიდევ უფრო განამტკიცოს კავშირი ან უკვე ცნობილ ვირუსურ ინფექციებსა და ვირუსური ეტიოლოგიის სიმსივნეებს შორის. ასევე გამოავლინოს საფუძვრები მომავალი კვლევებისათვის სიმსივნეთა ვირუსოლოგიის დარგში ჯერ კიდევ უცნობი ვირუსებისა და ვირუსური ეტიოლოგიის სიმსივნეებისათვის.

დიდი ხანია არსებობს ეჭვი რომ დედის ვირუსური ინფექციები ფეხმძიმობის დროს დაკავშირებულია ნაყოფის ბავშთა რეიკემიით დაავადებასთან. თუმცა ჯერჯერობით ვერ მოხერხდა ვერცერთი ცნობილი ვირუსის დაკავშირება ამ დაავადების ეტიოლოგიასთან. ჩვენ ვცადეთ გამოგვევლინა საფუძვრები მომავალი კვლევებისათვის ამ სფეროში და გამოვიყენეთ ახალი თაობის გენომური ანალიზატორები რათა გვეჩვენა თუ რომელი ვირუსების აღმოჩენაა შესაძლებელი დედის პლაზმიდან, რომელიც აღებუდ იქნა რეიკემიით დაავადებული ბავშვის ფეხმძიმობის დროს. დადგინდა რომ “ტორკუე ტენო ვირუსები” იყვნენ ყველაზე დიდი რაოდენობით წარმედგინილი დედათა პლაზმაში, ასევე ჩვენ აღმოვაჩინეთ “ტორკუე ტენო ვირუსის” ახალი ტიპები.

“მერკერის უჯრედების პოლიომავირუსი” აღმოჩენილ იქნა “მერკერის უჯრედების კარცინომას” ბიოლოგიურ ნიმუშებში. “მერკერის უჯრედების კარცინომა” წარმოადგეს კანის საკმაოდ იშვიათ და ავთვისებიან სიმსივნეს, რომელიც გაოირჩევა მაღალი სიკვდილიანობით. რათა გამოგვევლინა თუ კიდევ რომელ სიმსივნეებთანაა დაკავშირებული “მერკერის უჯრედების პოლიომავირუსი” ჩვენ გამოვიკვდიეთ კიდევ სხვა სიმსივნეების რისკი “მერკერის უჯრედების კარცინომით” დაავადებულ პაციენტებში. აღნიშნული მიზნისთვის გამოყენებულ იქნა სკანდინავიური ქვეყნების სიმსივნეთა რეგისტრები. დადგინდა რომ “მერკერის უჯრედების კარცინომით” დაავადებული პაციენტები გამოირჩევიან კანის სხვა სიმსივნეების მაღალი რისკით ჩვეულებრივ სკანდინავიურ პოპულაციასთან შედარებით. შესაძლო ფაქტორებიდან, რომლებმაც შეიძლება ახსნან ზემოთ აღნიშნული რისკი შეიძლება გამოიყოს ურტრა იისერი რადიაცია და “მერკერის უჯრედების პოლიომავირუსი”.

არსებობს ეჭვი რომ კანის “პაპილომა ვირუსები” დაკავშირებული არიან კანის სიმსივნეების განვითარებასთან. რათა გამოგვევლინა კავშირი ამ ორს შორის ჩვენ ჩავატარეთ მეტა-ანალიზი და სისტემატიური მიმოხილვა სტატიებისა რომლებიც იკვლევდნენ “პაპილომა ვირუსების” არსებობას კანის

სიმსივნეებით დაავადებულ პაციენტებსა და მათ ჯანმრთელ საკონტროლო ჯგუფში. ჩვენ დავაგინეთ რომ ბეტა-1, ბეტა-2, ბეტა-3 და გამა-1 “პაპილომა ვირუსები” განსაკუთრებით გავცეცდებურია კანის სიმსივნეებით დაავადებულ პაციენტებში მათ ჯანმრთელ საკონტროლო ჯგუფთან შედარებით.

ჩვენ ასევე გამოვიკვდიეთ კავშირი საშვიდსონოს ყელის სიმსივნის 8 სხვადასხვა ღონის დიაგნოზსა და 47 გენიტალურ “პაპილომა ვირუსს” შორის, რომელთაგანაც 12 კლასიფიცირებულია როგორც კარცინოგენური.

“პაპილომა ვირუსების” მრავალფეროვნება კანის სიმსივნით დაავადებული პაციენტებისგან აღებულ ბიოლოგიურ ნიმუშებში გამოვიკვდიეთ ახალი თაობის გენომური ანალიზატორებით. მეთოდის მგძნობელობის გამზრდისათვის ანალიზამდე ღნმ-ი ამპლიფიცირებულ იქნა PCR რეაქციით. კონვეციური PCR რეაქციის ნაკრია რომ მას შეუძლია აღმოაჩინოს მხოლოდ განსაზღვრული “პაპილომა ვირუსები”. ისინი რომლებიც ახლოს არიან PCR რეაქციის დროს გამოყვებენური პრაიმერ ღნმ-ბთან. ამის გამო არსებობს ეჭვი რომ გაცილებით მეტი “პაპილომა ვირუსი” არსებობს ვიდრე აქამდე აღმოჩენილ იქნა კონვეციური PCR რეაქციის მეშვეობით. ჩვენ ასევე გამოვიკვდიეთ თუ რომელი “პაპილომა ვირუსების” დაადგენაა შესაძლებელი ახალი თაობის გენომური ანალიზატორებით კანის სიმსივნისა და გენიტალური კონდილომის ნიმუშებიდან, ყოველგვარი წინასწარი PCR რეაქციის გარეშე.

ახალი თაობის გენომური ანალიზატორებით ჩვენ აღმოვაჩინეთ “პაპილომა ვირუსის” რამოდენიმე ახალი ტიპი კანის სიმსივნისა და გენიტალური კონდილომის ნიმუშებიდან, რომლებიც ვერ დაიჭირა კონვეციურმა PCR რეაქციამ. ეს შედეგები კიდევ ერთხელ მიუთითებს ახალი თაობის გენომური ანალიზატორების უპირატესობას კონვენციურ მოლკუდურ მეთოდებთან შედარებით.

წარმოდენილ თეზისში ჩვენ გამოვიყენეთ პოპულაციებზე დაფუძნებული სკანდინავიური რეგისტრები და ბიო-ბანკები, მეტა-ანალიზი და თანამედროვე ახალი თაობის გენომური ანალიზატორები რათა კიდევ უფრო გაგვეღრმავებინა არსებული ცოდნა სიმსივნეებთან დაკავშირებული ვირუსების ეპიდემიოლოგიის შესახებ და წარმოგვედგინა საფუძვლები მომავალი კვლევების გაგრძელების შესახებ ამ სფეროში.

LIST OF PUBLICATIONS

This thesis is based on the following papers, referred to in the text by their Roman numerals:

- I. BZHALAVA D, Bray F, Storm H, Dillner J. Risk of second cancers after the diagnosis of Merkel cell carcinoma in Scandinavia. *Br J Cancer*, 2011; 104:178-180
- II. BZHALAVA D, Ekström J, Lysholm F, Hultin E, Faust H, Persson B, Lehtinen M, de Villiers EM, Dillner J. Phylogenetically diverse TT virus viremia among pregnant women. *Virology* 2012;432:427-434
- III. BZHALAVA D, Johansson H, Ekström J, Faust H, Möller B, Eklund C, Nordin P, Stenquist B, Paoli J, Persson B, Forslund O, Dillner J. Unbiased approach for virus detection in skin lesions. *PLoS One*, 2013;8(6):e65953
- IV. BZHALAVA D, Guan P, Franceschi S, Dillner J, Clifford G. Systematic review of the prevalence of mucosal and cutaneous human papillomavirus types. *Virology*. 2013, doi:pii: S0042-6822(13)00435-2
- V. Ekström J, BZHALAVA D, Svenback D, Forslund O, Dillner J. High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. *Int J Cancer* 2011;129:2643-2650
- VI. Johansson H, BZHALAVA D, Ekström J, Hultin E, Dillner J, Forslund O. Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. *Virology*. 2013; 440:1-7

TABLE OF CONTENTS

1	INTRODUCTION.....	1
1.1	VIRUSES AND CANCER	1
1.1.1	<i>Tumor viruses</i>	1
1.1.2	<i>Human papillomaviruses</i>	2
1.1.3	<i>Anelloviruses</i>	5
1.1.4	<i>Cervical cancer and infections</i>	9
1.1.5	<i>Skin cancer and infections</i>	10
1.1.6	<i>Childhood leukemia and infections</i>	11
1.2	METHODOLOGIES FOR RESEARCH IN TUMOR VIRUS EPIDEMIOLOGY	12
1.2.1	<i>Registry linkage studies</i>	12
1.2.2	<i>Next generation sequencing and metagenomics</i>	15
1.2.3	<i>Meta-analysis</i>	20
2	PRESENT INVESTIGATIONS.....	22
2.1	AIMS	22
2.2	MATERIALS AND METHODS.....	23
2.2.1	<i>Patient data and bio-specimens</i>	23
2.2.2	<i>Methodologies</i>	23
2.3	RESULTS AND DISCUSSION.....	26
2.3.1	<i>Epidemiology of tumor viruses</i>	26
2.3.2	<i>High-throughput NGS technologies in the research on tumor virus epidemiology</i>	33
2.4	CONCLUDING REMARKS AND FUTURE PERSPECTIVES	41
3	ACKNOWLEDGEMENTS	43
4	REFERENCES.....	45

LIST OF ABBREVIATIONS

AK	Actinic keratosis
ALL	Acute lymphoblastic leukemia
ASCUS	Atypical squamous cells of undetermined significance
BCC	Basal cell carcinoma
CIN	Cervical intraepithelial neoplasia
CIN1	Mild cervical intraepithelial neoplasia
CIN2	Moderate cervical intraepithelial neoplasia
CIN3	Severe cervical intraepithelial neoplasia
CIS	Carcinoma in situ
EBV	Epstein–Barr virus
GAAS	Genome relative Abundance and Average Size
GRAMMy	Genome Relative Abundance estimates based on Mixture Model theory
GASiC	Genome Abundance Similarity Correction
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HPV	Human papillomavirus
HTLV-1	Human T-cell lymphotropic virus
HSIL	High-grade squamous intraepithelial lesion
HPyV	Human polyomavirus
HIV-1	Human immunodeficiency virus type-1
HR	High risk
IARC	International Agency for Research on Cancer
ICC	Invasive cervical cancer
KA	Keratoacanthoma
LSIL	Low-grade squamous intraepithelial lesion
LR	Low risk
MCV	Merkel cell polyomavirus
MCC	Merkel-cell carcinoma
NGS	Next generation sequencing
NMSC	Non-melanoma skin cancer
OTR	Organ transplant recipients

SCC	Squamous cell carcinoma
SIL	Squamous intraepithelial lesion
SIR	Standardized incidence ratio
TTV	Torque Teno virus
TTMV	Torque Teno-like Mini Virus
TTMDV	Torque Teno-like Midi Virus
PIC	Personal identity code
WGA	Whole genome amplification

1 INTRODUCTION

1.1 VIRUSES AND CANCER

1.1.1 Tumor viruses

Viruses were first suspected to be involved in tumor etiology almost a century ago when Rous [1] demonstrated that a solid tumor was transmissible to healthy chicken using cell free extract from tumor tissue [1]. Nowadays there are six established human tumor viruses: Epstein–Barr virus (EBV), Kaposi’s sarcoma herpes virus (HHV-8), hepatitis B virus (HBV), hepatitis C virus (HCV), human papilloma virus (HPV) and human T-cell lymphotropic virus (HTLV-1) [2,3]. The recently identified Merkel cell polyomavirus (MCV) [4] is classified as a probable carcinogen in the development of Merkel-cell carcinoma (MCC) [2,5]. The human immunodeficiency virus type-1 (HIV-1) is also classified as an established cancer-causing agent [2,3]. However, HIV-1 is not directly involved in the cellular transformation but is increasing the risk of cancer by causing immunosuppression [3].

In the year 2008, the International Agency for Research on Cancer (IARC) estimated that 16% of all new cancer cases worldwide (about two million annual cases) were attributable to infections [6]. About 65% of this number was attributable to viral infections such as HPV (30%), EBV (5.4%), HBV and HCV (29.5%) [6]. However, these figures might represent an underestimation of the true association [7] as the measurement of infection prevalence in the general population and/or in cancer patients is often inaccurate, which would tend to reduce the magnitude of the associations [7]. Even though only a minority of cancers are caused by viruses, the establishment of this association has resulted in large improvements in cancer control by virus-specific treatments and/or vaccination (e.g., against HBV and HPV) [8].

The last few decades have not only established that a considerable proportion of cancers are caused by viruses. They have also provided epidemiological indications that additional cancer-associated viruses may exist. Specific examples are the cancer forms that are increased among immunosuppressed individuals [9-15], as well as the space and time clustering of childhood leukemias [16]. The study of new relationships between virus infections and cancer faces several challenges: (i) Several oncogenic viruses are widespread and cause cancers only in a minority of infected individuals [3,7]; (ii) Furthermore, all cancers have a multifactorial etiology and their development almost always requires additional factors such as genetic alterations and/or immunosuppression. Thus, most of the human cancer associated viruses act as factors that initiate or promote the oncogenesis [3]; (iii) The incubation time before cancer development after initial infection with oncogenic virus might be several decades, making prospective studies difficult [3].

To investigate possible links between viruses and cancer a valid epidemiologic approach is necessary. An epidemiologic study is considered to be valid, when its design and methods are sound and provide a true estimate of the parameter of interest

[17]. To have true and unbiased estimates it is necessary to control all the factors, so called co-factors and/or confounders, that might be related to either exposure or outcome of interest (in this case to virus exposure or cancer development, respectively). To do this, the use of prospective studies nested in population based cohorts, such as biobanks with large study populations and long follow-ups are recommended [18]. Controlling confounding in a valid epidemiologic study requires meticulous planning of study design, such as how study subjects are selected, as well as how measurements of the values of interesting risk factors and other variables are going to be performed [17].

Another major challenge in cancer virology is the limitation of conventional molecular detection methods. Most studies in tumor microbiology have generally only studied one candidate infection at a time. However, different viruses and their occurrences may share several characteristics, which can act as confounding factors and may lead to biased epidemiologic results and/or inferences. Thus, to perform valid and unbiased epidemiologic studies on the association of viruses to cancer a measurement of as many of the microbes that are present as possible is necessary. Modern Next Generation Sequencing (NGS) technologies offer an opportunity to study potentially oncogenic viruses in the context of the entire microbiological community. A first and most important basis for further studies is therefore to provide a broad description of as many as possible of the known and unknown viruses that are present in relevant samples taken before cancer diagnosis. As the detection technology is powerful, it is likely that additional preventable cancer-associated viruses will be identified in the near future.

1.1.2 Human papillomaviruses

HPVs are small non-enveloped double-stranded DNA viruses that belong to the *Papillomaviridae* family. HPVs are a large and diverse group of viruses with 182 completely characterized types (www.hpvcenter.se), with new HPV types being continuously found [19-23].

Classification of HPVs is based on the nucleotide sequence of the capsid protein L1. HPV types belonging to different genera have less than 60% similarity within the L1 part of the genome. Different viral species within a genus share between 60 and 70 % similarity. A novel HPV type has less than 90% similarity to any other HPV type [24]. Novel HPV types are given a number only after the whole genome has been cloned and deposited with the International HPV Reference Center [24,25].

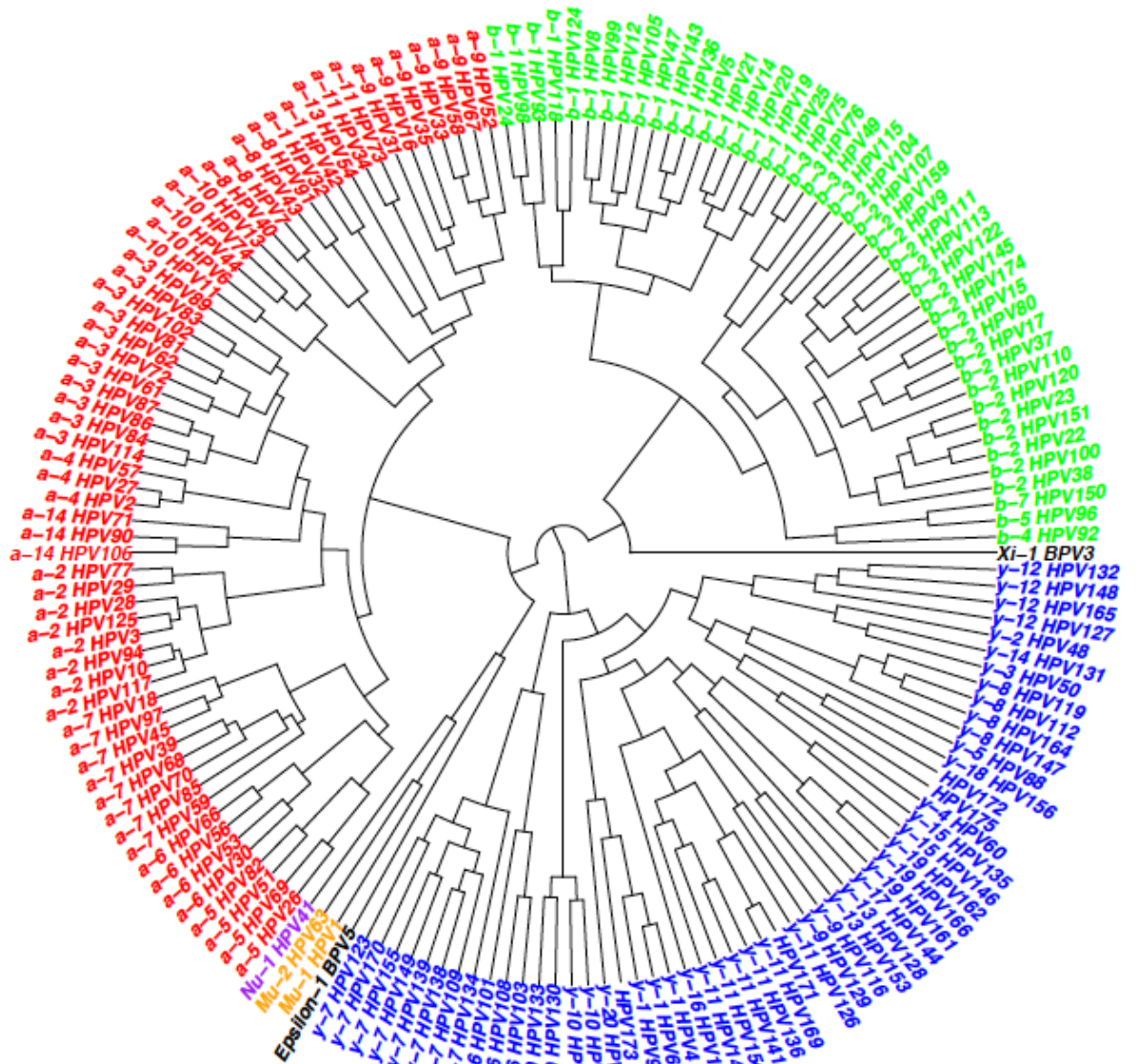


Figure 1. Phylogenetic tree of 164 HPV types and bovine papillomavirus type 3 and type 5. Alpha-, Beta-, Gamma-, Mu and Nu papillomaviruses are presented in red, green, blue, orange and purple colors, respectively. The phylogenetic tree is based on the L1 part of the genome.

Five major HPV genera are known: Alphapapillomavirus, Betapapillomavirus, Gammapapillomavirus, Mupapillomavirus and Nupapillomavirus [26] (figure 1).

The HPV genome contains three regions and approximately eight open reading frames (ORFs): (i) an early region with up to six ORFs (E1, E2, E4, E5, E6 and E7); (ii) the late region with two ORFs (L1 and L2); and the non-coding long control region [26] (figure 2).

Even though HPVs have a highly conserved structure of the genome, there are some differences among HPV types of different genera. Genomes of the alpha HPV types are relatively longer compared with the beta and gamma HPVs. Also, the E5 ORF is missing from the genomes of most of the beta- and gamma HPVs. At least three gamma HPV types (HPV101, HPV103, and HPV108) also lack the E6 ORF. Beta HPV

types encode a longer E2 ORF compared to HPVs from other genera. It is not clear how these differences affect the lifecycles of these viruses.

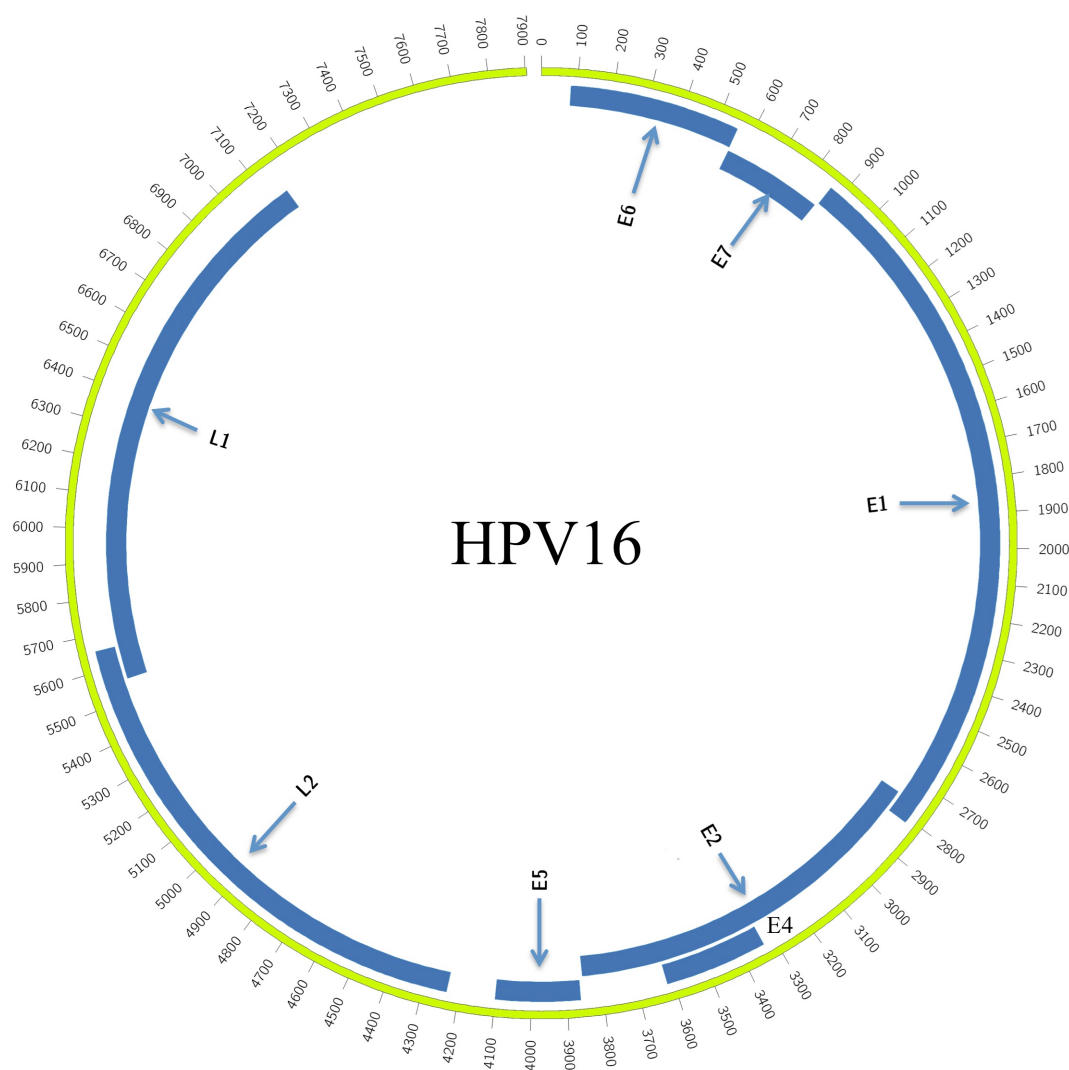


Figure 2. Genomic organisation of the HPV16 genome. Light blue lines represent protein coding ORFs of the HPV16 genome. Plots were generated using Circos visualization tool [27].

HPVs are epitheliotropic and infect cutaneous or mucosal stratified epithelia of humans [3,26] and cause a wide range of diseases from benign lesions to invasive tumors [28,29]. The majority of sexually active women will have a genital HPV infection at least at some point in their lives and genital HPVs are considered as one of the most common sexually transmitted diseases [26]. Number of sexual partners, absence of circumcision among males and cigarette smoking [30-33] are the known risk factors for an HPV infection. Immunosuppressed patients, such as organ transplant recipients and/or HIV positive individuals, have an increased prevalence of both single and multiple HPV infections, compared with the healthy population [34]. Usually, genital HPV infections, are asymptomatic and transient and are cleared within two years after the initial infection in the majority of women [30]. However, persistent infection will develop in approximately 10% of infected women [3]. A persistent HPV infection is a prerequisite for the development of cervical cancer [3].

The oncogenic mucosal HPV types in the alphapapillomavirus genus are a major cause of cervical cancer, HPV16 and HPV18 being the most frequent types [29,35,36]. They are also linked to the development of vulvar, vaginal, anal, penile and oropharyngeal cancers [3,37,38].

Mucosal HPVs are classified as high- and low-risk types depending upon their degree of carcinogenicity [2]. In 2009, the IARC working group classified 12 mucosal HPV types (HPV16, HPV18, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58 and HPV59) as established to be carcinogenic to humans (Group 1) [2,39], sometimes referred to as high-risk (HR) HPV types. These 12 types cluster together in the same evolutionary branch or "high-risk clade" that includes alphapapillomavirus species groups 5, 6, 7, 9 and 11. Other types in the high-risk clade were classified as possible carcinogens (Group 2B) based upon their phylogenetic relatedness to established (Group 1) types, with the exception of HPV68, which was classified as a probable carcinogen (Group 2A) based on some, but limited, epidemiological evidence. There are also benign mucosal HPV types in the alphapapillomavirus genus for example HPV6 and HPV11, that cause benign genital warts (condylomas) [25].

There are 45 recognized HPV types in the beta genus (www.hpvcenter.se). Only a limited number of functional and epidemiological studies has investigated their oncogenic potential. Studies based on in vitro and in vivo experiments indicated oncogenic properties of several beta HPV types, such as HPV5, HPV8 and HPV38 [40-45]. Epidemiologic studies also noted a link between detectability of antibodies against beta HPV and/or their DNA and non-melanoma skin cancers (NMSCs) [46-49]. However, the studies were inconsistent and a systematic review on the association of beta HPVs with cutaneous lesions has been lacking [3].

The gamma genus includes 54 recognized HPV types (www.hpvcenter.se). They are highly prevalent on the skin of the general population and little information is available about their biological properties. Although some findings suggest an association with NMSC [50], even a systematic review of the literature has identified only a limited number of observations [51] and further research is necessary to clarify if they have a carcinogenic potential.

1.1.3 Anelloviruses

In 1997, the widespread anelloviruses were discovered. They form a large and diverse group of non-enveloped, single-stranded DNA viruses with a circular, negative-sense genome ranging in size from 2 to 3.8 kb [52]. Three anelloviruses, able to infect humans, are classified into Alphatorquevirus (Torque teno virus (TTV)), Betatorquevirus (Torque Teno-like Mini Virus (TTMV)), and Gammatorquevirus (Torque Teno-like Midi Virus (TTMDV)) genera of the *Anelloviridae* family of viruses [53] (figure 3).

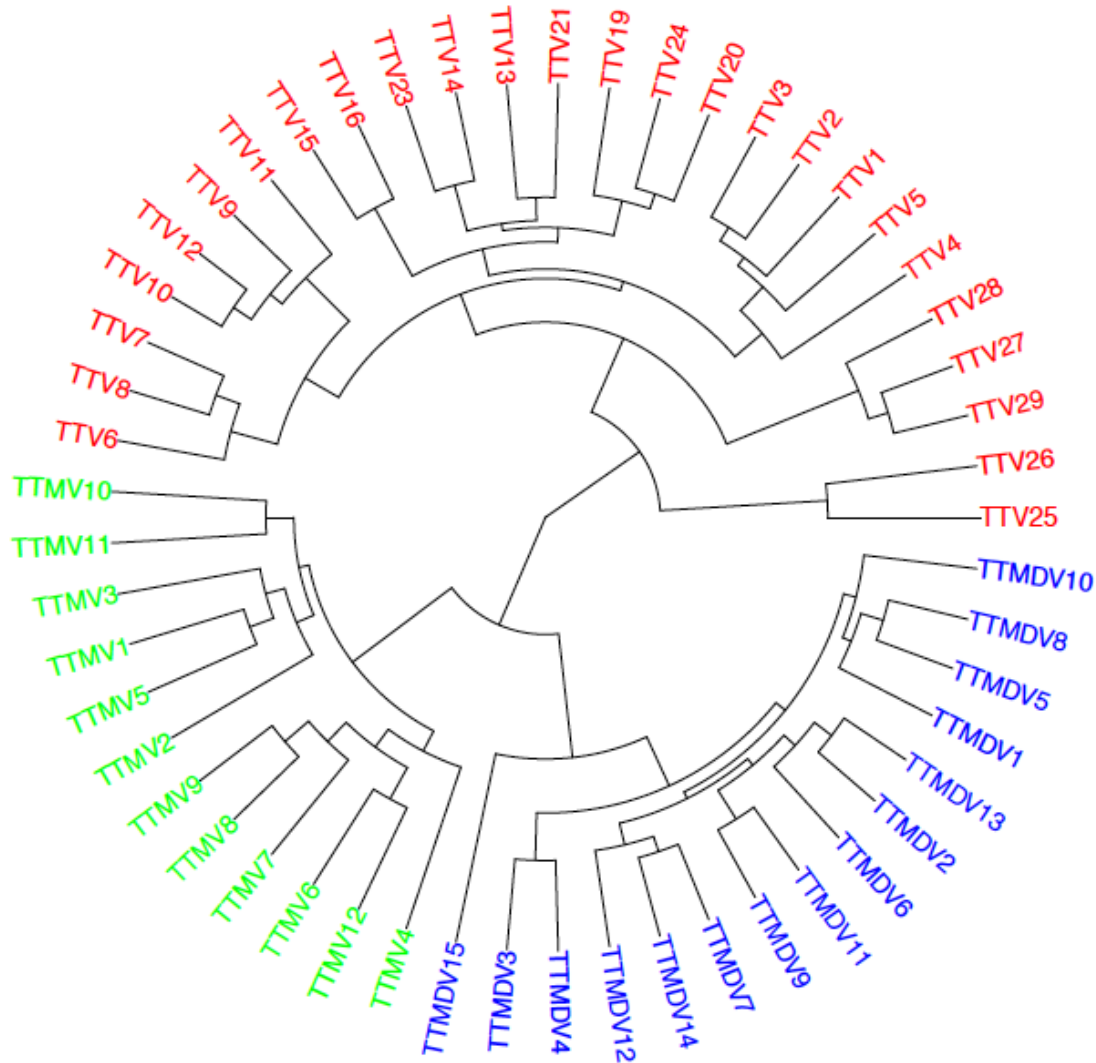


Figure 3. Phylogenetic tree of the *Anellovirus* family. Alpha-, Beta-, Gamma Anelloviruses are presented in red, green and blue colours, respectively.

Anelloviruses show extreme diversity both within and between species [53,54]. On the nucleotide level they can exhibit as much as 33%–50% divergence [53,54]. Although there is an extreme genetic diversity, the members of the *Anelloviridae* family are also characterised by a conserved genomic organization. Their genomes consist of two main ORFs (ORF1 and ORF2), as well as several additional smaller ORFs resulting from splicing events and a non-coding GC rich region [53,54] (figure 4). Anelloviruses also share several conserved protein signatures such as an arginine-rich N-terminus in ORF1 [55,56]; four binding sites for Rep proteins involved in rolling circle replication (two of which were reported to be conserved among many plant and animal Circoviruses) [55,56]; the protein motif W-X7-H-X3-C-X1-C-X5-H in ORF2, which was reported to be common for TTV, TTMVs and chicken anemia virus [57,58]; a serine-rich domain in the C-terminal region of ORF3; and the E-X8-R-X2-R-X4-6-P-X5-11-P-X1-8-V-X1-F-X1-L motif in the C-terminal region of ORF4 [59].

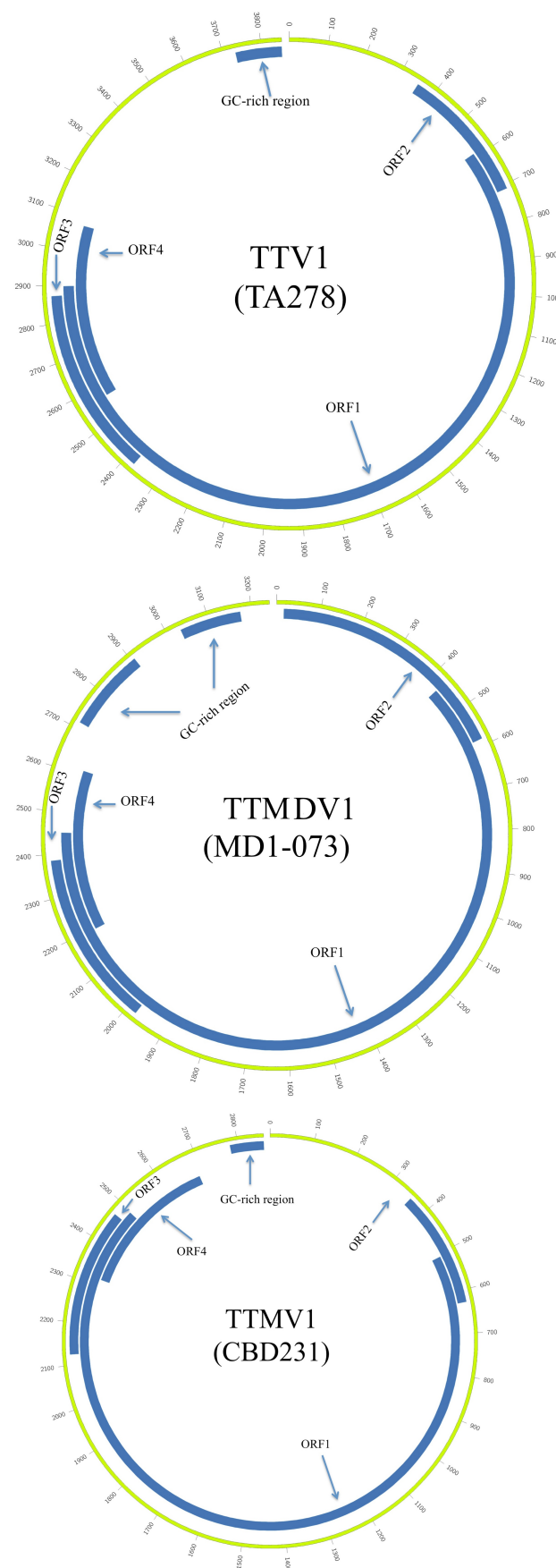


Figure 4. Genomic organisation of TTV1, TTMDV1 and TTMV1 genomes. Plots were generated using Circos visualization tool [27].

Anellovirus infections are highly prevalent in the general population [60]. Presence of their DNA has been found in nearly every organ, tissue and body fluid of humans tested [52]. TTV has an ability to sustain a lifelong viremia even in healthy individuals [61].

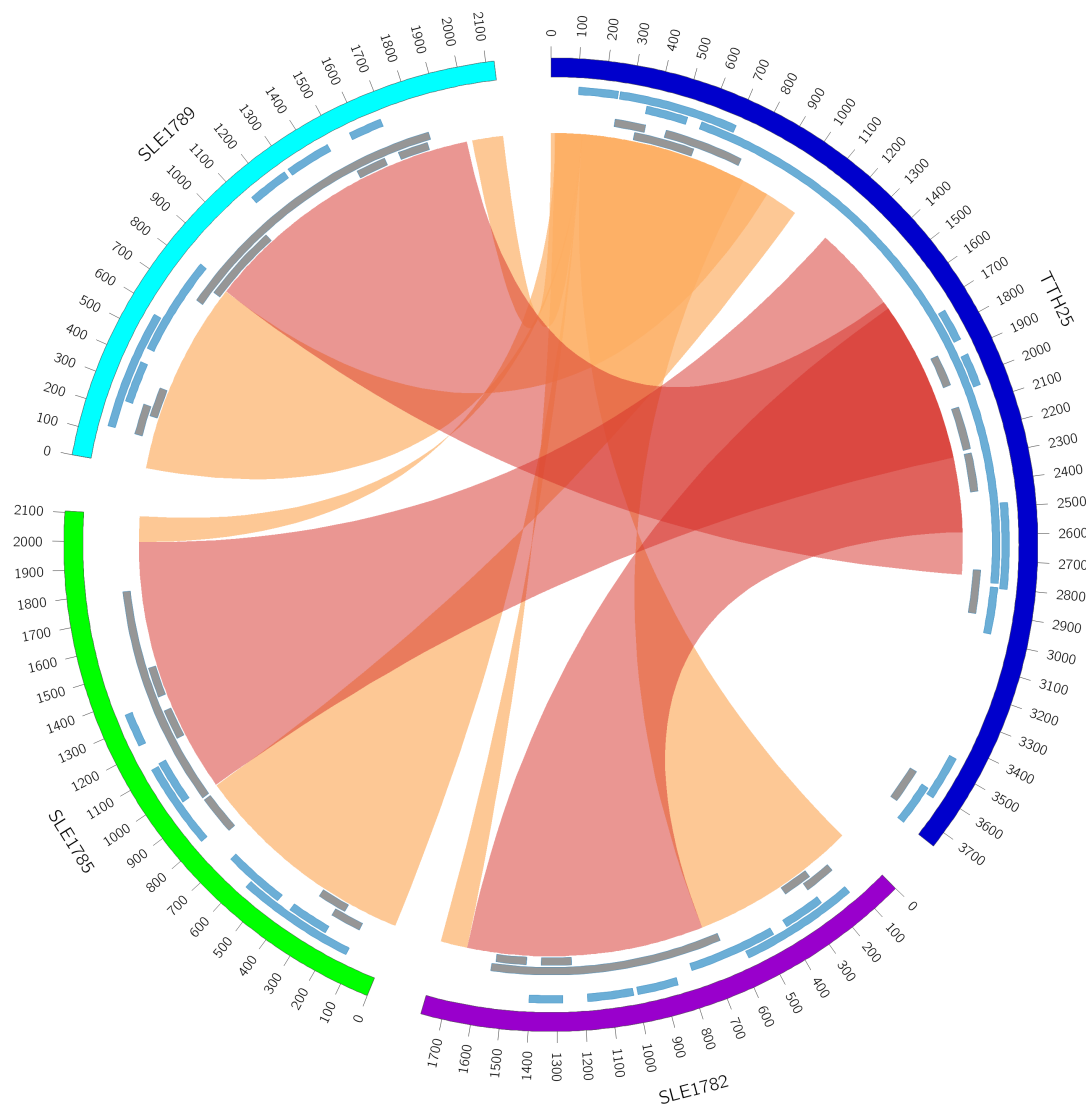


Figure 5. Intragenomic rearrangements between complete genome tth25 (dark blue) and closely related subviral genomes of sle1782 (purple), sle1785 (green) and sle1789 (turquoise) isolated from serum samples from mothers to leukemic children [62]. Orange and red links between genomes represent fragments of tth25 genome arranged on subviral molecules on either in sense or antisense, respectively. Light blue and grey lines represent sense and antisense ORFs, respectively. Plots were generated using Circos visualization tool [27].

Anelloviruses have been studied in the context of many diseases. However, because of their nearly universal presence and persistent viremia in human populations, investigations to link the TT viruses to the etiology of specific diseases have yielded inconsistent results and no direct link has been established [63-65]. It has been suggested that certain genotypes or groups of genotypes of anelloviruses may be

pathogenic [61,62]. In vitro analysis of replication and transcriptional activities of full length TTV genome has provided evidence for the origin and selection of the smaller subviral molecules through intra-genomic rearrangements [61,62] (figure 5). This raises a possibility that such infections in an individual over time may lead to formation of pathogenic strains through this phenomenon. This possibility has received some support from in vitro studies on the transforming effect of TT virus fusion proteins generated by viral recombination [61].

Because of the extreme diversity of anneviruses and their potential to rearrange genomes, a comprehensive and unbiased analysis of the viral DNA, present in a sample, is necessary to investigate their possible link to human diseases. Also it is crucial to study them in the context of their community. This can only be achieved using metagenomic analysis based on next generation massively parallel sequencing technologies.

1.1.4 Cervical cancer and infections

Cervical cancer is the second most common cancer among women worldwide with a majority of the cases (83%) occurring in developing countries [66].

Non-invasive precancerous lesions are divided into different grades of cervical intraepithelial neoplasia (CIN) or squamous intraepithelial lesions (SIL). Based on the degree of cytological atypia of epithelial cells, CIN is graded as mild dysplasia (CIN1), moderate dysplasia (CIN2) or severe dysplasia (CIN3)/ carcinoma in situ (CIS). In the Bethesda classification system, CIN1 corresponds to low-grade SIL (LSIL) and CIN2-3 correspond to high-grade SIL (HSIL) [67]. The term atypical squamous cells of undetermined significance (ASCUS) is used to describe poorly visualized cells from LSIL or HSIL [68].

Persistent infections with one or more HR HPV types are the major cause of cervical cancer [69], with HPV16 and HPV18 being the most important [35,70]. Besides HR HPV types, several other co-factors, such as smoking [71], multiparity [72] and sexual behaviours (e.g. age at first intercourse and lifetime number of sexual partners [73]) also contribute to the increased risk of cervical cancer. Other sexually transmitted infections such as herpes simplex virus type 2 [74] and Chlamydia trachomatis [75] have also been reported as co-factors. However, the possibility exists that the association seen with other STIs may be due to confounding by HPV (their presence could be an indication of a higher risk behaviour that increases the exposure to HPV). In prospective studies, only Chlamydia trachomatis has consistently been found to associate with cervical cancer [76].

1.1.5 Skin cancer and infections

The two major forms of NMSC, squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), of the skin are two of the most prevalent cancers among Caucasian populations worldwide. BCC is approximately four times as common as SCC [77].

Most of the new NMSC cases occur in patients that are over 60 years of age [78]. NMSCs (excluding BCC) represent the second most common cancers in both sexes and are the most rapidly increasing tumors in Swedish population [78]. Over the last decade an average 4.9% and 7.3% annual increase was observed for men and women, respectively, in Sweden [78].

MCC is a rare and aggressive neuroendocrine malignancy of the skin [79]. The majority of MCC cases occur in Caucasian populations [80]. Incidence rates for MCC are extremely age-dependent and most cases occur in patients older than 65 years [80]. Even though the incidence of new cases remains low, it is increasing annually [80].

Solid organ transplant recipients (OTR) have an approximately a 65- to 100-fold increase of SCC [10-13]; a 2- to 16-fold increase of BCC [14,15] and 10-fold increase of MCC incidence [81] compared to the general population, suggesting that the development of these cancers is under control of the immune system that is being suppressed in OTR.

The elevated risk in immune-compromised individuals has suggested that the immune system may target a viral antigen expressed in precancerous cells, in turn suggesting that an infection may be involved in the etiology of NMSC [10-13]. HPV has been the most commonly studied candidate infectious agent [46,82]. An association of HPV infection with skin cancer was first demonstrated in patients with the rare hereditary disease epidermodysplasia verruciformis [83]. These immunosuppressed patients are highly susceptible to HPV infections, that often progress to SCC [83]. The cutaneous HPV types are commonly found in skin lesions, often as multiple infections with many different HPV types. Studied lesions include benign skin warts [84], actinic keratoses (AKs), NMSCs [37] and keratoacanthomas (KAs) [46,85]. Epidemiological studies have found that betapapillomaviruses are more frequently detected in SCC patients than in their healthy controls [82,86-89]. Seropositivity for beta HPV antibodies has been reported to be associated with SCC [50,90]. Several studies also demonstrated that prevalence of these HPVs is higher in AK than in SCC [91,92]. AK is a precursor of SCC and this observation might indicate involvement of betapapillomaviruses in the early stages of carcinogenesis. This effect is not reported for BCC [47]. Healthy individuals are also frequently positive for betapapillomavirus DNA [93,94] and it seems that these tend to persist on healthy skin more often than HPV from other genera [95].

Ultraviolet (UV) radiation is a well-known risk factor for NMSCs, as these cancers are most often found on areas of the skin that are regularly exposed to sunlight or other UV radiation [10,80,96]. Several studies have demonstrated that E6 and E7 proteins from

HPV5, HPV8 and HPV38 may contribute to UV-induced carcinogenesis by inhibiting DNA repair mechanisms [40-45]. This observation indicated that HPVs from the beta genus might act as co-factors and facilitate the accumulation of UV-mediated mutations.

Gammapapillomaviruses are also suspected to be involved in SCC carcinogenesis [50]. However, available data is inconsistent and based on quite small number of observations [51] and further research is necessary. Inconsistencies between studies could be attributable to small samples sizes and the extreme diversity of gamma HPV types, that could conceivably lead to misclassification of the HPV types present, with the detection methods that have been used so far [51]. Widespread infections with multiple of HPV types present at low viral loads have made it difficult to perform reliable epidemiologic studies. However, NGS holds promise in being able to provide a more reliable HPV typing, as the nucleotide sequence of the viruses present in samples is obtained.

Human polyomaviruses (HPyV) are the second largest group of viruses that are also implicated to be associated with the skin cancers. However, similar to HPVs, they are also diverse, widespread and also found on healthy human skin [97]. HPyV6, HPyV7, and MCHPyV are the most commonly found polyomaviruses on human skin [97,98].

MCV was identified in MCC [4] and it is the only HPyV that is classified as a probable carcinogen. A majority of MCC tumors are positive for MCC DNA [99,100]. However, the majority of healthy adults have antibodies against MCV [101-103]. MCV DNA is also present in skin swabs, skin biopsies, and plucked eyebrow hairs of healthy subjects [97,104]. The fact that the MCV genome is clonally integrated in MCC tumor cells [4] supports the possibility that MCV is involved in the development of MCC. Also, MCC tumor cells tend to have higher viral loads of MCV DNA compared to other MCV DNA positive tissues [99,100]. A prospective epidemiological study nested in population based biobanks found that presence of MCV antibodies was associated with an increased risk for future MCC [105].

Metagenomic analysis of different skin lesions, using next generation massively parallel sequencing technologies found that a majority of the viral sequences originated from different HPVs. Some HPyVs and anelloviruses were also found [19,21]. Future studies using NGS are needed to provide epidemiological-scale data on whether any one of these viruses are associated with any human disease.

1.1.6 Childhood leukemia and infections

Leukemias, the most common childhood cancers in the developed world, are biologically diverse clonal diseases originating from single blood cell progenitors that have accumulated mutations [106,107]. The etiology of childhood acute lymphoblastic leukemias (ALL) is not known [107]. Studies of Guthrie cards and identical twins with concordant leukemia have provided strong evidence that the existence of translocations

giving rise to fusion genes is usually of fetal origin [107]. However, similar translocations are also present in healthy individuals and monozygotic twins may develop leukemia at different ages [107], indicating the need of an additional event to develop an overt disease [107]. Reports of indirect epidemiological characteristics such as observed protective effects of intermittent infections during the first year of life, attendance at whole day care during the same period and a marked inverse risk with birth order and sibship size have argued for a possible infectious etiology of ALL [108]. However, no specific agent(s) have been identified and possible mechanisms for involvement of infections in leukemogenesis is unclear [107-109].

zur Hausen and de Villiers postulated that initial events of leukemogenesis is triggered by a pre- or perinatal infection [108]. Continuous production and proliferation of infected cells, would lead to high load of the suspected infectious agent [108], which may be a risk factor for the development of full-blown leukemia. Viral load and thus leukemia risk would be decreased if the immune system produces interferon as a result of intermittent infections after birth. According to the model, a putative leukemogenic agent synergistically co-operates with in-utero or perinatally acquired chromosomal rearrangements [107].

Studies attempting to investigate the risk of ALL after infections during pregnancy [110-126] and after delivery [16,109,127] is scarce, controversial and have not identified any specific infectious agent. Maternal infection with EBV [114,116] and neonatal adenovirus-C infection [125] was reported to be associated with development of childhood leukemia in the offspring. However, follow-up studies could not confirm this [117,126]. As conventional technology used for analyses suffers from low-throughput and from being inherently biased in detecting only sequences with homology to the PCR primers used, further progress in this area will most likely require use of NGS technology.

1.2 METHODOLOGIES FOR RESEARCH IN TUMOR VIRUS EPIDEMIOLOGY

In this section, major modern day methodologies for research in tumor virus epidemiology will be discussed, in particular (i) registry-linkage studies (ii) high-throughput NGS technologies and (iii) systematic review and meta-analysis. These methods have all been used in papers included in this thesis.

1.2.1 Registry linkage studies

In the Nordic countries, a series of high quality population-based biological specimen banks and patient data registries exist, with many decades of follow up [128]. Different biobanks and data registries are possible to link using the unique personal identity code (PIC), providing unique possibilities to conduct longitudinal molecular epidemiological research with adequate statistical power even for rare diseases and exposures [128]. In

this section I will discuss Nordic biobanks and data registers which were used in the studies included in this thesis.

1.2.1.1 Maternity cohorts

The Finnish Maternity Cohort, at the National Public Health Institute, contains about 1.5 million serum samples from more than 98% of all pregnant women (approximately 1 million) in Finland. They were collected at maternity care units at 12–14 weeks of gestation, for the purpose of screening of congenital infections since 1983 and onwards and currently include 7 million person years of follow up [128].

The Icelandic Maternity Cohort contains 98 000 serum samples from more than 95% of all pregnant women (approximately 48 000) in Iceland. The samples have been collected at maternity care units at 12–14 weeks of gestation, for the purpose of screening of congenital infections. Samples are stored in the centralized Department of Medical Virology, Landspítali University Hospital since 1980 and onwards and include 600 000 person years of follow up [128].

The Northern Sweden Maternity Cohort contains approximately 120 000 serum samples from 86 000 pregnant women in Northern Sweden collected at maternity care units during 14 week of gestation, for the purpose of screening of congenital infections. Samples are stored at the virus laboratory of Umeå University since 1975 and onwards. The cohort has 1.2 million person years of follow up [128].

The Southern Sweden Maternity Cohort, which contains approximately 100 000 samples from 74 000 pregnant women collected at maternity care units at 14 week of gestation for screening of virus infections and rubella immunity and stored at the Skåne Biobank [128]. The cohort has been collected consecutively since 1989 and now has 750 000 years of follow up.

All samples have been stored at -20°C to -25° C. The corresponding databases contain the PIC, enabling linkage to nationwide cancer registries [128].

1.2.1.2 Cancer registries

Population-based and countrywide Nordic cancer registries were established almost 50 years ago and they are notified of virtually all histologically confirmed new cases of cancer [129]. Nordic cancer registries are considered to have a consistently high degree of comparability and completeness overtime [129].

1.2.1.3 Case-control identification by registry-linkages

In epidemiologic research, registry-linkage means connecting data for a particular individual across different data items (e.g. between data files of cancer registries and biospecimen banks) to identify data about exposure and outcome of interest [18].

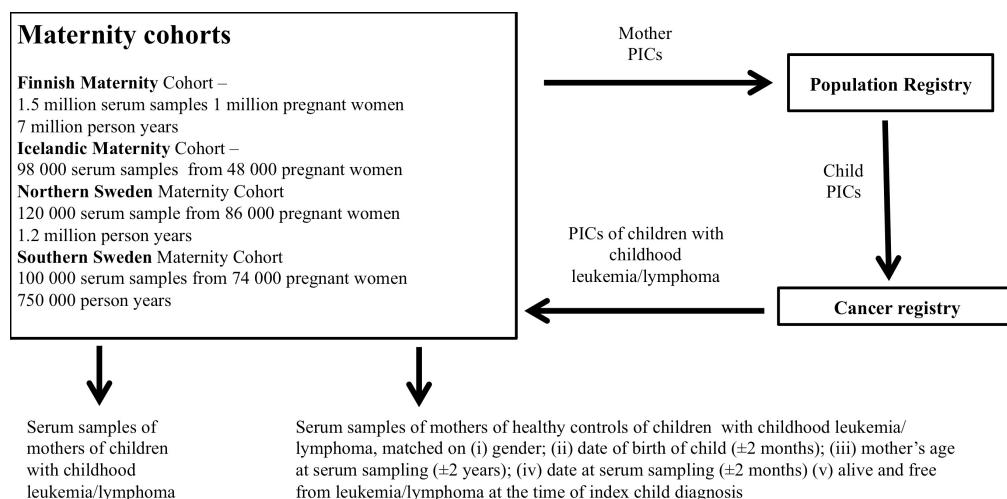


Figure 6. Registry-linkage pipeline to identify serum samples of mothers of children who developed childhood/leukemia lymphoma and their healthy controls.

Registry-linkage between Nordic cancer registers and bio-specimen banks was conducted to enable the study of maternal infections during pregnancy and risk of childhood leukemia in the offspring [114,116,130]. To investigate the role of infections during pregnancy and risk of childhood leukemias in the offspring it is necessary to (i) identify children which developed childhood leukemia/lymphoma and (ii) their mothers which donated biological samples to biobanks during their pregnancy, also called the index pregnancy. In Sweden, registry-linkage is conducted through the following steps: PICs from all female donors in the biobank are sent to the tax office authorities. Tax office authorities use the population registry to identify all children born to a woman who has a sample in the biobank. Then PICs of all these children is sent to cancer registry to identify who developed childhood leukemia/lymphoma. The file from the cancer registry with PICs of childhood leukemia/lymphoma cases is sent back to the biobank and is linked to the file received from tax office authorities to get the mothers PICs. Identified mothers PICs are linked to the microbiology biobank to get all available samples and the blood draw date of the index mothers (figure 6).

After identifying index pregnancy samples the next step is to identify samples of healthy controls (figure 6). To control for possible confounding factors, control subjects are selected per case matched for (i) gender; (ii) date of birth of child (± 2 months); (iii) mother's age at serum sampling (± 2 years); (iv) date at serum sampling (± 2 months) (v) alive and free from leukemia/lymphoma at the time of index child diagnosis. If desired number of matched control subjects per case can not be found, the matching criteria for age of child (ii), mother's age at serum sampling (iii) and sample storage time (iv) can

be widened stepwise by one month for child's age, one year for mothers age and one month for sample storage; If there are more than desired number of matched control subjects, controls may be randomly selected within eligible subjects. The reasons that these factors have been chosen for matching are the following: gender (i) and age (ii) are known cofactors for many diseases, including childhood leukemia/lymphoma [107]; mother's age (iii) might be related to risk of development of childhood leukemia/lymphoma. Including date at serum sampling (iv) in the matching parameters gives us the opportunity to improve comparability of measurements from frozen biological material where some substances of interest may decay as a function of the length of storage [131]. Also, it allows us to control for seasonality (e.g. during the winter time there is a higher chance to be infected with influenza virus, compared to summer time). Finally for a valid comparison we need matched controls that were alive and free from the disease of interest at the time the matched case was diagnosed with the disease (v).

This example illustrates the unique opportunity that Nordic population based cancer registries and bio-banks can provide. Molecular epidemiological studies of a rare disease, such as childhood leukemia/lymphomas could be conducted and viral exposures investigated when case subjects and their matched, healthy controls were still in their fetal development [116]. A total of 343 serum samples of mothers of childhood leukemia/lymphoma cases and their 943 healthy controls were tested for EBV antibodies [116]. Maternal EBV reactivation was statistically significantly related to childhood leukemia/lymphoma development in the offspring [116]. However, this was not confirmed in a later study [117]. One of the possible explanations for this could be that we are unable to control for confounding factors (e.g. infection with other viruses) due to pitfalls of conventional molecular diagnostic methods.

1.2.2 Next generation sequencing and metagenomics

1.2.2.1 Viral metagenomics

The term human microbiome or microbiota, defines the collection of microorganisms that reside in the human body [132]. The viral fraction of human microbiome is referred to as the human virome [133,134]. Viruses constitute only a small part of human microbiota, but their proportion and composition seems to change in diseased individuals [135,136].

Viruses can be found in every human individual and the number of orphan viruses (viruses which are not linked to any diseases) is continuously increasing [134]. For example, for most of the beta- and gammapapillomaviruses there is yet no direct evidence regarding association to human skin diseases [93,94]. Lazarczyk et al speculated that betapapillomaviruses might even establish a symbiotic relationship with humans under certain conditions and actively participate in the keratinocyte proliferation during wound healing [137]. Another example comes from hepatitis G virus, which is also not clearly linked to human diseases [138]. Survival in HIV-

infected individuals was associated with hepatitis G virus [138]. However, further research is necessary to elucidate whether there exists any possible symbiotic roles of viruses.

Studies on viral metagenomics require unbiased sequencing of all DNA in biospecimens, in contrast to studies of bacterial communities where the conserved 16S rDNA typically is targeted. Viruses usually have smaller genomes than other microbes and virus-related sequences will usually constitute only a small fraction of all sequences.

NGS technologies can be used to obtain a comprehensive and unbiased sequence of the DNA present in a sample, without the need of any prior PCR or other amplification that requires prior information about sequences that may be present [139].

The complete sequencing of all microbiological sequences that may be present in a sample is termed metagenomics [140]. Viral metagenomics is nowadays routinely used for virus detection and is commonly used for discovery of new viruses [4,19-22,130,141-143]. Viral metagenomics provides an opportunity to perform a large-scale analysis of all infections that are present in cancers and in healthy individuals. Thus, it has the potential to further our knowledge of the role of viruses in human diseases such as cancer. Sequencing of cancer specimens with NGS has already been used in the discovery of a new cancer-associated virus (namely MCV) [144].

1.2.2.2 Next generation sequencing instruments

During the past decade, there has been a dramatic evolution of NGS instruments like 454 GS FLX (Roche), SOLiD (ABI), Ion Torrent Proton (Life Technologies) and Genome Analyzer/HiSeq System (Illumina). A variety of bench-top NGS instruments have also been developed, e.g. the 454 GS Junior (Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies) and are now becoming a standard equipment in virological laboratories. However, NGS instruments generate huge amounts of data and to analyse this data is one of the biggest challenges in the use of NGS for viral diagnostics and research.

1.2.2.3 Bioinformatics for viral metagenomics

The bioinformatics pipelines to analyse NGS data usually start by quality checking according to their Phred quality scores [145] (Figure 7). 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 a base calling accuracy of 99% (1 error per 100bp) [145]. Specific quality filtering conditions can be adapted for different downstream analyses [146].

NGS technologies might produce exact and/or nearly duplicated reads due to errors in PCR amplification and/or sequencing errors [147,148]. Presence of duplicated reads might also introduce an overestimation of the species abundance. On the other hand, duplicated reads might also include natural duplicates that by chance originate from the same start from the same genomic position [147,148]. Highly abundant species have a higher chance to have natural duplicates [148] and their removal might introduce bias towards underestimation of abundances [147]. To decrease sampling variation and discard redundant data, sequence datasets are normalized using a digital normalization algorithm (<http://ged.msu.edu/papers/2012-diginorm>). Normalized datasets have substantially less size and require significantly reduced computational resources for *de novo* assembly.

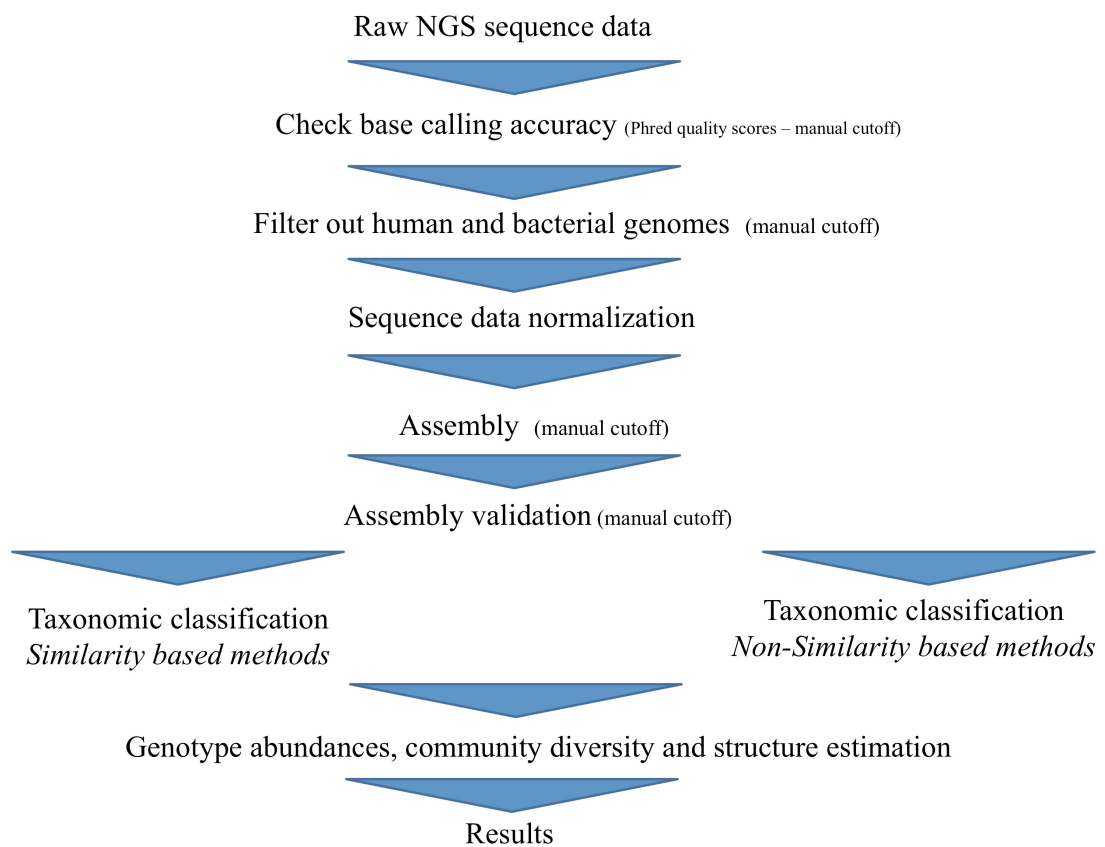


Figure 7. Bioinformatics pipeline to analyse high-throughput sequencing data for viral metagenomics.

NGS data from human samples subjected to whole genome amplification (WGA) typically contain more than 70% of human-related sequences. Unless there has been prior separation of viral capsids or shorter DNAs from long chromosomal DNA [19] (Table 1), viral reads typically constitute less than 1% of the reads (Table 1). With prior selection for viral nucleic acids, the human and bacterial related reads will still be the most commonly obtained reads, followed by sequences classified as “other” and “unknown” [19,130] (Table 1). Enrichment for viral particles by ultracentrifugation is helpful in the analysis of serum samples but has not been useful in the analysis of biopsies or skin swabs (Table 1). Bacterial sequences and sequences classified as “other” and “unknown” may also be present in negative control samples (water) after

NGS [130] (Table 1), and it is therefore imperative that all metagenomic sequencing projects also include sequencing of negative control samples [130]. The background sequences found in water samples might be present due to the background reactivity of the Phi29 polymerase reaction [149] or represent environmental contamination. However, water controls have so far been found to be uniformly negative for viral sequences [130]. To obtain the dataset that contains reads of interest, e.g. the virus-related reads for viral metagenomics, sequences that are not a target of the investigation need to be filtered out on the bioinformatics level. This will further speed up downstream analysis and decrease the risk of mis-assemblies [146].

Table 1. Typical taxonomic assignment of NGS reads (%). Summary of results in previous studies using different types of biospecimens, pre-treatments and NGS platforms.

Sample type	FFPE ¹ biopsies		Fresh frozen biopsies		Skin swabs				Serum		Water
Pre-amplification treatment after WGA	E-gel	-	E-gel	-	E-gel	UC ²	-	-	UC ²	-	-
Sequencing platform	GS FLX	GS FLX	GS FLX	GS FLX	GS FLX	GS FLX	GS FLX	Ion PGM 300bp kit	GS FLX	Illumina MiSeq	GS FLX
Human	63.9	37.3	95.5	99.8	42.6	2.1	69.1	77.3	81.5	75	2.8
Bacteria	14.6	21.3	3.1	0.1	36.8	61.1	24.2	18.3	6.9	1	52.2
Virus	0.0	0.2	0.0	0.0	0.1	0.1	0.3	0.4	0.4	0.1	0.0
Other	11.7	10.2	0.5	0.0	17.1	31.5	2.2	1.3	8.6	0.5	15.5
Unknown	9.8	30.9	0.9	0.0	3.5	6.1	4.2	2.7	2.7	24.4	29.5

¹Formalin Fixed Paraffin Embedded. ²Ultracentrifugation.

NGS technologies produce billions of short reads from random locations in the genome by oversampling it. Assembly algorithms, in the process called *de novo* assembly, reconstruct original genomes 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 (OLC) assemblers, most widely applied to the longer reads and *de Bruijn Graph* Assemblers, most widely applied to the shorter reads. To validate assembly results, several assembly algorithms are used, as well as re-mapping of all singletons reads to assembled contigs [19,139].

The possibility always exists that assembly algorithms may construct erroneous “chimeric” sequences by the assembly of two different 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 algorithm to identify possible “chimeric” HPV sequences [20]. It is based on the assumption that an HPV genome should have similar degree of identity over its entire genome to the most closely related HPV type. 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). This is checked by the

following procedure: the sequence that aligns to its most closely related sequence in GenBank is divided into three equal segments. If at least one of the segments had less than 90% similarity and at least one more than 90% similarity, as well as if the difference between these segments by similarities to corresponding overlapping parts 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”. However, this approach can’t be used for anelloviruses, as they frequently rearrange parts of their genomes with each other (figures 5). For anelloviruses, we assessed the level of assembly coverage, protein coding potential and the conserved protein signatures [130].

One of the biggest challenges for bioinformatics analysis is taxonomic classification of NGS data as many of the sequences have no homologs in the public databases or are highly divergent, which is especially true for viral sequences [150]. Taxonomic classification of metagenomic reads can be divided into similarity and non-similarity-based methods. One of the most famous similarity-based taxonomic classifications is performed by NCBI BLAST searches, where sequences are compared to known genomes. However, a large part of the sequencing reads from *de novo* sequencing projects are classified as unknown [19,130]. This can result from incompleteness of public sequence databases or drawbacks of NGS technologies such as short read lengths and sequencing errors. Because metagenomes might contain a large amount of sequences that have very distant homologs or even no homologs at all in public databases, more sensitive algorithms, such as BLASTx and tBLASTx searches are conducted against the protein database after the BLASTn search on the nucleotide level.

To classify sequences from alignment results, several methods have been developed. One of the first and most frequently used is MEGAN [151]. In BLAST searches, sequences might have multiple matches and MEGAN finds the ‘Lowest Common Ancestor’ node of all matching sequences in the phylogenetic tree, which reduces the risk of false positive matches. However, MEGAN might produce false negative results by discarding sequences if they do not satisfy user-defined cut-offs. Because the size of genome is related to the number of reads in metagenomic samples, MEGAN is suboptimal for quantitative metagenomic analyses. This problem has been addressed by the development of the GAAS (Genome relative Abundance and Average Size) tool [152] that iteratively weights each reference genome for all matching reads and the number of reads is then normalized to the length of their genomes. GRAMMy (Genome Relative Abundance estimates based on Mixture Model theory) [153] is another useful tool that, compared to GAAS models, reads assignment ambiguities, genome size biases and read distributions along the genomes on a unified probabilistic framework [153]. However, both GAAS and GRAMMy estimate similarities from the alignment qualities of the reads to the reference genomes and not from the reference genomes directly. Thus, 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 [154].

1.2.3 Meta-analysis

Meta-analysis is defined as analysing multiple studies and combining their results using statistical methods [155,156].

Some individual studies have small sample sizes resulting in that they would not be able to detect a true effect if there is one [155]. This is frequently a problem in research on cancer and viruses. Correctly planned and conducted meta-analyses provide an opportunity to improve the precision by increasing the statistical power, as well as to answer questions that could not be answered by individual studies [155]. Also it gives an opportunity to investigate whether the diversity of study methods influence the outcomes, if there is consistency of the effect estimates from different studies, as well as if there is publication bias or gaps in the literature [155,156]. Greenland and O'Rourke [156] defined meta-analysis as the “study of studies and their results” [156]. It can also be useful to generate new hypothesis by investigating questions that were not asked by the original studies [155]. However, meta-analysis has also a risk to produce seriously misleading results. For example, if critical parts of individual studies have within-study biases, there will be variations across studies. If this is not carefully considered, meta-analysis can assemble these errors and interpret them to be valid [155,157]. In the data analysis step meta-regression methods such as fixed-effect models and random effect models are usually employed to control for between-study heterogeneities [158].

The fixed effect model assumes that there is no between-study heterogeneity in the treatment effect and the observed variability is due to chance differences created from the study population (sampling variability) [156,158]. The random-effects model assumes that both different treatment effects and sampling variability are accountable for the variability of observed estimates [156,158]. For example, different studies might use molecular detection methods with different sensitivity; and/or the study population in one study is more susceptible to disease than in another study. Thus, according to the random effects model, a true effect could be variable between studies even with large study populations and statistical powers [156,158].

If there were no heterogeneity between studies both fixed and random effect models would give identical results. With high between-study heterogeneity, the random effect model will produce wider confidence intervals compared to the fixed-effect method [156,158].

Meta-analysis can be improved with diagnostic procedures such as sensitivity and influential analysis. In sensitivity analysis, the summary estimate is examined if it is sensitive to any particular factor and/or assumption [156,158]. Thus meta-analysis will be conducted twice, one with factor and/or assumption and one without and if inference is insensitive there will be little change between the two [156,158]. Influential analysis examines how a particular study or group of studies affects the inference of the meta-

analysis. It will be conducted several times with or without the studies of interest [156,158].

Before drawing final conclusions from meta-analysis, publication bias should be also investigated. Studies with null or non-significant results have a lower chance to be published than studies with positive and significant results. Thus, publication bias may influence the inference in a meta-analysis study.

As only statistical tools are unable to deal with heterogeneity among studies, it is extremely important to account for study differences already in the data retrieval step [156]. For example, different study designs (such as retrospective case-control and prospective cohort designs), as well as different analysis methods, different comparators and their combinations could be analysed separately [155].

2 PRESENT INVESTIGATIONS

2.1 AIMS

Paper I: to estimate the excess risk of cancers among patients with MCV-associated disease, by investigating the risk of secondary cancers after the diagnosis of MCC in Denmark, Norway and Sweden.

Paper II: To comprehensively analyse which known and unknown viruses are present in serum samples of pregnant women, using high-throughput next generation sequencing technology.

Paper III: To investigate the presence of virus DNA in skin lesions, such as SCCs, AKs, and KAs using several next generation sequencing technologies.

Paper IV: To investigate the prevalence of different types of HPVs across a broad range of disease grades, to gain basic knowledge of how widespread infections with the different HPV types are, as well as to provide information on the possible carcinogenicity of different HPV types.

Paper V: To achieve an improved resolution of the diversity of HPV types in lesions such as SCCs, AKs and KAs, using next generation sequencing technology.

Paper VI: To investigate the presence of virus communities in condylomas which were apparently “negative for HPV” by conventional PCR methods using unbiased next generation sequencing technology.

2.2 MATERIALS AND METHODS

2.2.1 Patient data and bio-specimens

In **Paper I** the risk of secondary cancers after the diagnosis of MCC was investigated using the population based national cancer registries in Denmark, Norway and Sweden. All persons who had been registered with MCC over the calendar period of 1980–2007 in Denmark and during 1990–2007 in Norway and Sweden were included in the study cohort. The follow-up started on the date of the diagnosis of MCC and ended on the date of death, emigration or the closing date of the study (31 December 2007), whichever occurred first. Only cancers that occurred at least 6 months after an MCC diagnosis were considered. Secondary MCC following primary MCC, were excluded from the analysis.

In **Paper II** serum samples of 112 mothers to leukemic children were identified through registry linkages of cancer registries and maternity cohorts in Finland, Iceland and Sweden.

For **Paper III** and **Paper V** skin biopsies were collected from immunocompetent patients with lesions diagnosed as SCC (n=86) or AK (n=92), attending Swedish and Austrian hospitals, as well as biopsies from 92 KAs from both immunosuppressed and immunocompetent patients at the Department of Dermatology and Plastic Surgery at the Norwegian National Hospital, Oslo, Norway.

Paper IV is a systematic review and meta-analysis of 47 mucosal HPV types in cervical samples across the entire range of cervical diagnoses from normal to cervical cancer. The analysis was restricted to studies using a number of well-characterized PCR assays. For the cutaneous HPV types, meta-analysis was restricted to studies that assayed their prevalence in skin diseases in a case-control format.

For **Paper VI**, 42 "HPV-negative" condyloma swab samples were collected from 21 women and 19 men and were analysed using unbiased next generation sequencing.

2.2.2 Methodologies

In **Paper I**, the observed subsequent primary cancers, occurring at least 6 months or at least 1 year after MCC diagnosis, were compared with those expected from the incidence rates among the national populations. Nordic Cancer Registry (NORDCAN) incidence rates specific for country, age, gender, 5-year calendar period and site were multiplied by the respective numbers of accumulated person-years at risk to estimate the expected number of cancers. The ratios of the observed-to expected number of cases were expressed as the standardized incidence ratio (SIR). Risks were also estimated, stratified by the time that had passed since diagnosis of the first cancer (<1-2 year, 1–4.9 years, ≥ 5 years). Ninety-five percent confidence intervals of SIR were computed assuming a Poisson distribution for observed cases. Findings were considered significant for two-sided $p < 0.05$.

In **Paper IV**, the systematic review and meta-analysis, studies of interest were identified through PubMed searches for studies published up to 12th of May of 2013. Data retrieval and analysis was performed separately for mucosal and cutaneous HPV types.

For mucosal HPV types, combinations of search terms such as “cervical cancer”, “cervical intraepithelial neoplasia”, “HPV”, “human,” “female” and “polymerase chain reaction” were used. Eligible studies needed to: (i) use broad-spectrum consensus PCR assays based on the primers MY09/11, PGMY09/11, GP5+/6+, SPF10, SPF1/GP6+ or L1C1/L1C2, and (ii) report overall and type-specific HPV prevalence by strata of cytopathological and/or histopathological cervical diagnoses. Cases were classified into eight grades of cervical diagnosis: those diagnosed by cytology as (i) normal; (ii) ASCUS; (iii) LSIL or (iv) HSIL; those diagnosed by histology as (v) CIN1; (vi) CIN2 or (vii) CIN3 (including squamous carcinoma in situ) and those diagnosed as (viii) invasive cervical cancer (ICC), which comprises both squamous cell carcinoma, adeno/adenosquamous carcinoma or cervical cancer of other/unspecified histology. From the eligible studies, the following data were extracted by cervical diagnosis: country, source of HPV DNA (cells versus biopsies/tissue), PCR primers, sample size, as well as overall and type-specific prevalence of HPV DNA. Paper IV reported mucosal HPV DNA prevalence as a percentage of all women tested by consensus PCR. Each HPV type was evaluated independently and denominators thus vary by HPV type as many studies tested only for some of the mucosal HPV types.

As for cutaneous HPV types, a combination of search terms such as “Human papillomavirus”, “HPV”, “cutaneous” and “skin” was used to identify studies of interest. To be eligible, studies needed to meet the following criteria: (i) provide analyses on exposures to cutaneous HPV types and risk of non-melanoma skin cancers, recorded separately for SCC of the skin, BCC or for AK; (ii) compare healthy and diseased individuals and not merely reporting cutaneous HPV sequences. Studies reporting other case groups than skin cancers were excluded, as well as review papers. If several studies reported on the same population, the study that was most recently published was chosen. From the identified studies data was extracted on the number of diseased subjects and their healthy controls by cutaneous HPV type exposure status, by laboratory method of exposure identification and by types of skin cancers. The “Meta” package from the statistical software R (www.r-project.org) was used to estimate summary effect estimates. Both fixed- and random-effects models were used and we weighed each study by the inverse variance method.

In **Paper II**, **Paper III**, **Paper V** and **Paper VI** we employed high-throughput NGS technologies to investigate which known and yet unknown viruses were present in bio-specimens from patients with cancer or who later developed cancer. In **Paper II** extracted DNA from serum samples were ultracentrifuged followed by WGA using GenomiPhi High Yield. In **Paper III** three sample preparation methods were employed: E-gel followed by WGA, ultracentrifugation followed by WGA and direct

WGA of the sample. In **Paper VI** direct WGA of the samples was employed, while in **Paper V** we used E-gel followed by WGA.

In **Paper II**, WGA amplified DNA from serum samples of 112 mothers to leukemic children were pooled in three different pools: pool A –15 mothers from the Finnish maternity cohort; pool B - 78 mothers from the Finnish and 19 from the Swedish maternity cohorts; pool C - 22 mothers from the Finnish and Icelandic maternity cohorts. Pool C was also subjected to TTV amplification by general primer PCR. In **Paper III** and **Paper VI** WGA amplified DNA samples were pooled into seven and 10 different pools, respectively. In **Paper V** extracted DNA was amplified using the general HPV primers FAP and mixed to three different pools.

Pooled samples, as well as individual samples from **Paper V** were subjected to high-throughput sequencing on GSFLX 454 (Roche). The pool of swab samples of SCCs & AKs from **Paper III** was also sequenced on Ion Torrent PGM (Life Technologies) using Ion Torrent 300 and 400 bp sequencing kits.

NGS data from **Paper II**, **Paper III**, **Paper V** and **Paper VI** were analysed as described in the section of Bioinformatics for Viral Metagenomics, above.

2.3 RESULTS AND DISCUSSION

In this thesis, we performed a population based registry-linkage study (**Paper I**) to evaluate the excess risk of other cancers among patients with the MCV-associated cancer MCC to investigate if there may be shared etiology between MCC and other cancers. In several papers, we used high-throughput NGS technologies to identify which known and yet unknown viruses are present in biospecimens of patients with or due to develop cancer, clearly demonstrating the usefulness of high-throughput NGS technologies in research on tumor virus epidemiology (**Paper II**, **Paper III**, **Paper V** and **Paper VI**). Finally, a systematic review and meta-analysis was performed to investigate the prevalence of infections with different HPV types in various forms of cervical and cutaneous diseases as a proxy for possible carcinogenicity.

Results will be presented in two sections (1) epidemiology of tumor viruses and (2) high-throughput NGS technologies in the research on tumor virus epidemiology.

2.3.1 Epidemiology of tumor viruses

2.3.1.1 *Registry linkage study*

Using nationwide cancer registries from Denmark, Norway and Sweden, a total of 756 patients diagnosed with MCC were identified (**Paper I**). Seven hundred and sixteen of them were diagnosed with MCC as a first cancer. Overall cancer incidence was increased among the cohort of patients diagnosed with MCC compared with the general population in Denmark, Norway and Sweden (SIR 1.38 (95% confidence interval (CI): 1.10-1.72)) (Figure 8). As for specific cancer forms, patients diagnosed with MCC were at an excess risk for non-melanoma skin cancers (SIR 8.35 (95% CI: 5.97-11.68)), melanoma of skin (SIR 4.29 (95% CI: 1.93-9.56)) and laryngeal cancer (SIR 9.51 (95% CI: 2.38-38)) (Figure 8).

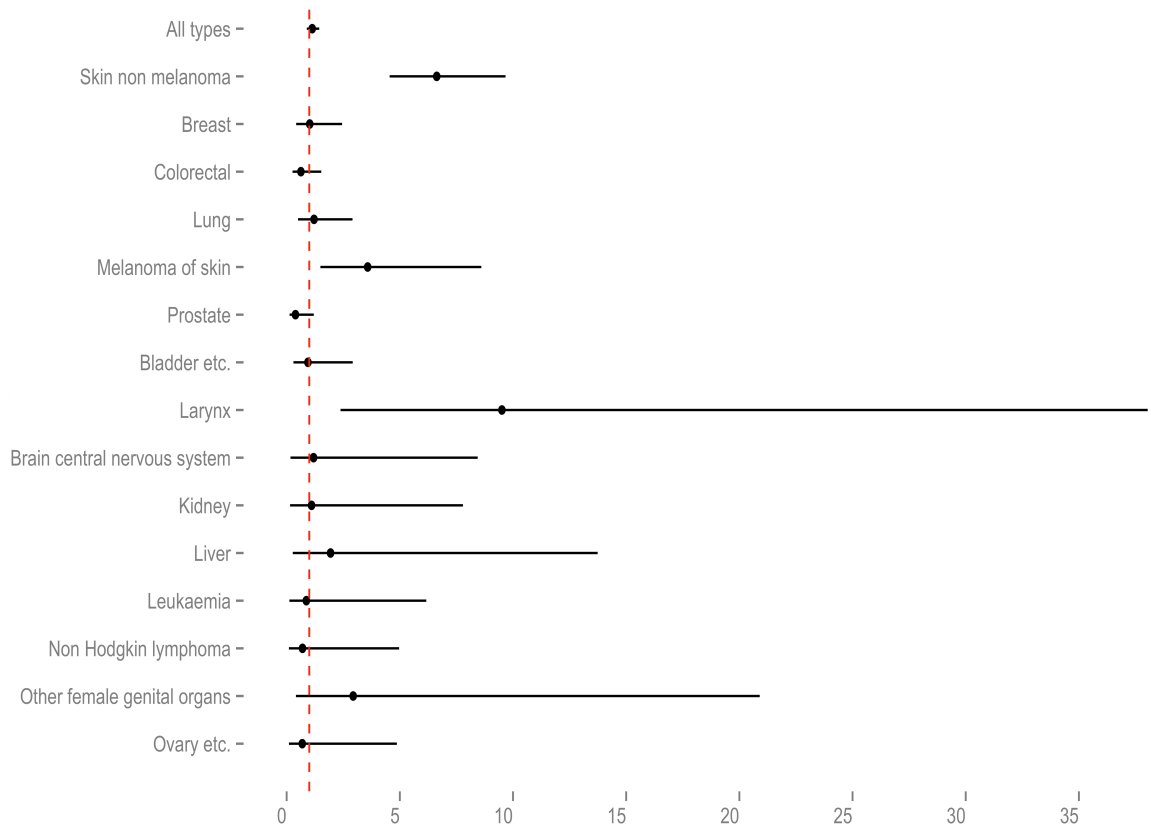


Figure 8. Forest plot of SIR estimates and their 95% CI of cancer incidence in the cohort of MCC patients, compared to general population. Black circles represent point estimates of SIR. Black lines represent 95% CI for respective SIR. If they cross the red dashed line, the result is not statistically significant.

The major strength of **Paper I** is that it is based on joint data from cancer registries of three Nordic countries. All MCCs are diagnosed histologically, which should provide a correct diagnosis. To assess the risk for second cancers requires a large study cohort and a long follow up time. MCC is a rare and aggressive cancer with poor prognosis and it occurs mostly in elderly patients. Thus, the numbers of person-years available for follow-up is limited. Even though we combined comprehensive data from three Nordic countries, the numbers of person years of follow up was still a limitation of **Paper I**. Surveillance bias could also be a possible limitation. However, to avoid this bias we excluded cases occurring less than six months after MCC diagnosis and also found limited changes in estimates by restricting the study to cancers occurring >1 year after diagnosis.

In conclusion, **Paper I** found that the incidence of second primary cancers is elevated among patients diagnosed with MCC compared to the general population. Patients diagnosed with MCC are at an excess risk in particular for NMSC, melanoma of the skin and larynx cancers (Figure 8).

The results in **Paper I** are only partially in line with results from two other registry-linkage studies from USA [159] and Finland [160]. Howard et al [159] reported elevated risks of cancers for salivary gland, brain, biliary sites, multiple myeloma,

chronic lymphocytic leukemia and Non-Hodgkins Lymphoma [159]. The Finnish Cancer Registry identified a significantly increased risk for BCC of the skin and for chronic lymphocytic leukemia after the diagnosis of MCC [160].

Possible explanations for the excess risk that we found could include factors such as the impact of increased surveillance of the skin for patients diagnosed with MCC (surveillance bias). Among possible shared causative factors, exposure to UV and/or infection with MCV can be considered.

The results of **Paper I** suggest that further studies on a possible link between MCV and NMSCs might be motivated.

2.3.1.2 Systematic review and meta-analysis

Our systematic review and meta-analysis (**Paper IV**) identified a total of 423 and 15 eligible studies for assessing prevalence of mucosal HPVs and exposure to cutaneous HPV types among patients with skin cancers, respectively.

2.3.1.2.1 Mucosal HPV types

Studies of mucosal HPV types assessed the prevalence of 47 different types among 371,951 women across eight grades of cervical diagnoses (table 2). Overall prevalence of HPVs increased with increasing severity of cervical disease from 12.6% in normal cytology to as high as 89.5% in ICC (table 2).

All HPV types classified as established or probably carcinogenic by IARC (Class 1/2A), were more commonly found among patients with ICC than among individuals with normal cytology. Higher prevalences of HPV16, HPV18 and HPV45 were detected in ICC than in any other grade of cervical lesion. This supports the carcinogenicity of these types. HPV16 was the most frequently detected type in every grade of cervical diagnosis. This could be due to an advantage of HPV16 over all other mucosal HPV types in terms of transmissibility and/or persistence. HPV16 seems more efficient to escape the host immune-surveillance compared to other HR HPV types [161].

HPV types from group 2A/2B (probably or possibly carcinogenic), such as HPV26, HPV67, HPV68, HPV69, HPV73 and HPV82 were also more commonly present in ICC than in normal cytology (table 2). Further research may eventually accumulate data to consider these types as established carcinogens and thus becoming targets for cervical cancer prevention.

Table 2. Type-specific prevalence of mucosal HPV DNA, by grade of cervical diagnosis (adapted from Paper IV).

HPV type	IARC classification	% of positive of tested samples							
		Normal	ASCUS	LSIL	HSIL	CIN1	CIN2	CIN3	ICC
Any		12.6	52.1	75.2	85.3	74.2	85.4	92.4	89.5
HPV16	1	2.6	12	19.5	40.5	19.2	34	53.8	55.8
HPV18	1	1	4.7	6.3	8.2	7	8.7	6.9	14.3
HPV45	1	0.6	2.9	3.3	3.9	3	4.3	3.4	4.8
HPV33	1	0.6	3	4.9	7.1	3.7	7.1	8.5	4.0
HPV58	1	0.8	3.9	5.5	6.9	7.1	10.3	8.4	4.0
HPV31	1	1	4.7	7.9	9.4	6.8	10	10.8	3.5
HPV52	1	1	5.4	6.5	8.6	10.1	14.1	9.6	3.2
HPV35	1	0.4	3	3.8	5	2.8	4.3	3.3	1.6
HPV39	1	0.6	4.2	5.5	3.8	4.9	4.7	3.1	1.3
HPV59	1	0.4	3.1	4.3	2.7	3.7	4	2.1	1.2
HPV51	1	0.9	4.8	9.4	6	8.1	8.4	5.1	1.0
HPV56	1	0.6	3.5	7	3.1	5.6	3.7	2.3	0.8
HPV68	2A	0.4	1.8	2.2	1.8	2.3	2.5	1.9	0.5
HPV53	2B	1.1	5.4	8.4	4	6.4	4.5	3.1	0.5
HPV73	2B	0.3	1.9	2.7	2.5	2.3	2.1	1.8	0.5
HPV6	-	0.8	3.1	5.9	2.7	6	3	1.7	0.4
HPV11	-	0.5	1.6	2.9	0.9	2.5	1.4	0.8	0.4
HPV62	-	1.0	4.4	4.1	3.3	6.1	4.1	2.3	0.4
HPV54	-	0.6	2.4	2.9	2.7	2.6	3.2	2	0.3
HPV66	2B	0.6	4	7.7	3.5	5.9	4.6	2.4	0.3
HPV67	2B	0.2	1.2	1.8	1.1	1.9	1.5	0.7	0.3
HPV84	-	0.5	2.8	3.1	3.2	3.3	3	1.8	0.3
HPV26	2B	0.1	0.5	0.4	0.5	0.6	1.1	0.5	0.2
HPV30	2B	0.1	0.5	0.3	0	0.6	0.4	0.5	0.2
HPV69	2B	0.1	0.2	0.3	0.3	0.3	0.4	0.4	0.2
HPV70	2B	0.8	2.4	2.3	2.1	1.5	1.7	1.1	0.2
HPV81	-	0.6	2.3	2.8	1.8	2.9	3.1	1.1	0.2
HPV82	2B	0.1	1.2	1.8	2	1.5	1.8	1.7	0.2
HPV34/64	2B	0.1	0.3	0.3	0.2	0.2	0	0.1	0.1
HPV42	-	0.5	4.3	4.8	1.6	3.3	2.4	1.2	0.1
HPV44	-	0.4	0.1	0.3	0.3	1.2	0.9	0.4	0.1
HPV55	-	0.3	1.2	2.2	1.8	1.4	1.4	1.2	0.1
HPV61	-	0.6	3.5	3.8	3.5	2.7	2.9	2.2	0.1
HPV71	-	0.4	0.7	0.4	0.3	0.3	0.4	0.4	0.1
HPV72	-	0.3	0.7	0.8	0.8	0.5	0.5	0.6	0.1
HPV83	-	0.4	2.1	1.9	1.8	2.2	1.4	1.5	0.1
HPV89	-	0.4	3	3.4	2.9	5.2	3.1	2.3	0.1
HPV90	-	0.6	1.9	1.9	0.7	0.8	0.8	0.9	0.1
HPV32	-	0.1	0.3	0.1	<0.1	0.1	0.6	0.1	<0.1
HPV40	-	0.2	0.9	1.6	0.6	0.8	1	0.4	<0.1
HPV43	-	0.3	0.5	0.3	0.1	0.4	0.2	0.1	<0.1
HPV57	-	<0.1	0.1	<0.1	<0.1	<0.1	<0.1	0.1	<0.1

Table 2 Continued from previous page

HPV type	IARC classification	Normal	ASCUS	LSIL	HSIL	CIN1	CIN2	CIN3	ICC
HPV74	-	0.5	0.2	0.6	0.8	0.5	0.7	0.1	<0.1
HPV85	2B	0.2	0.4	0.4	0.5	0.2	0.3	0.2	<0.1
HPV86	-	0.1	0.4	0.4	<0.1	<0.1	<0.1	<0.1	<0.1
HPV87	-	0.1	0.5	0.2	0.4	0.4	<0.1	0.5	<0.1
HPV91	-	0.2	3.6	4.2	3.1	2.5	2.6	2.5	<0.1
Multiple infection		4.3	21	28	28	32	39	27	12

Some HR HPV types, such as HPV31, HPV51 and HPV52, had higher prevalence in intermediate grades than in ICC (table 2), suggesting that they may cause these lesions, but that lesions caused by these HPV types may have a lower progression potential to progress to ICC.

A major finding of **Paper IV** is that a number of non-HR HPV types, such as alpha-3 types HPV61, HPV62, HPV84 and HPV89, were commonly detected in low and high-grade cervical abnormalities (table 2). Also, high prevalence of multiple HPV infections was noted in these lesions (table 2). This suggests that infections with multiple HPV types may be involved in the etiology of these low and high-grade cervical abnormalities. It is estimated that the proportional impact of HPV-16/18 vaccination on cervical lesions can be predicted to rise from 17% of ASCUS, through 49% of HSIL, up to 70% of ICC. However, these estimates are based on the assumption that HPV16 and HPV18 are causally related to the lesion in which they are found, and doesn't take into account the presence of other HPV types. Findings of **Paper IV** indicate that this assumption may lead to an over- or under-estimation of the proportional impact of individual types, particularly in low- and high-grade cervical abnormalities.

2.3.1.2.2 Cutaneous HPV types

The identified studies in **Paper IV** enabled an analysis on the association of different skin lesions and HPV types from genus beta, gamma, mu and nu (Figure 9).

For the beta genus, a large number of studies investigated their association with SCC with serology and/or HPV DNA detection methods. The individual type level data showed that the prevalence of antibodies against Beta-1: HPV8; Beta-2: HPV15, HPV17, HPV38, HPV49; and Beta-3: HPV76 were elevated in SCC patients in comparison to their controls (Figure 9). However, risk differences of these types using DNA detection did not reach statistical significance (Figure 9). The opposite effect was observed for the prevalence of HPV24 (Beta-1) and HPV92 (Beta-4). They were significantly higher in SCC patients than controls using DNA detection methods, but the corresponding difference in antibody prevalence did not meet statistical significance (Figure 9). A similar effect was observed between HPV24 (Beta-1) and AK, a SCC

precursor, prevalence of HPV24 DNA was significantly higher in cases than controls, but not with serology methods. When we aggregated type level data at the species level, the prevalence of Beta-1, Beta-2 and Beta-3 species were each statistically elevated in SCC patients compared to their healthy controls, both using serology and DNA detection methods (Figure 9).

Data on HPV types from the genus Gamma, Mu and Nu was scarce and was coming from studies that used serology. At an individual HPV type level, there were no significant differences between cases and controls (Figure 9). However, when type level data was aggregated on species level, Gamma-1 species antibody prevalence was significantly elevated in SCC compared to controls (Figure 9).

None of the cutaneous HPVs appeared to be significantly elevated in BCC in comparison to controls, neither by serology nor DNA detection methods (Figure 9).

In conclusion, findings of **Paper IV** about cutaneous HPV types showed that prevalence of species Beta-1, Beta-2, Beta-3 and Gamma-1 were each significantly elevated in SCC compared to their healthy controls. However, this effect was not observed among patients with BCC. On the HPV type level, cutaneous HPV types were frequently found to have non-significant tendencies for increased risk of SCC. This indicates that further studies on the presence of HPV types in SCC are warranted.

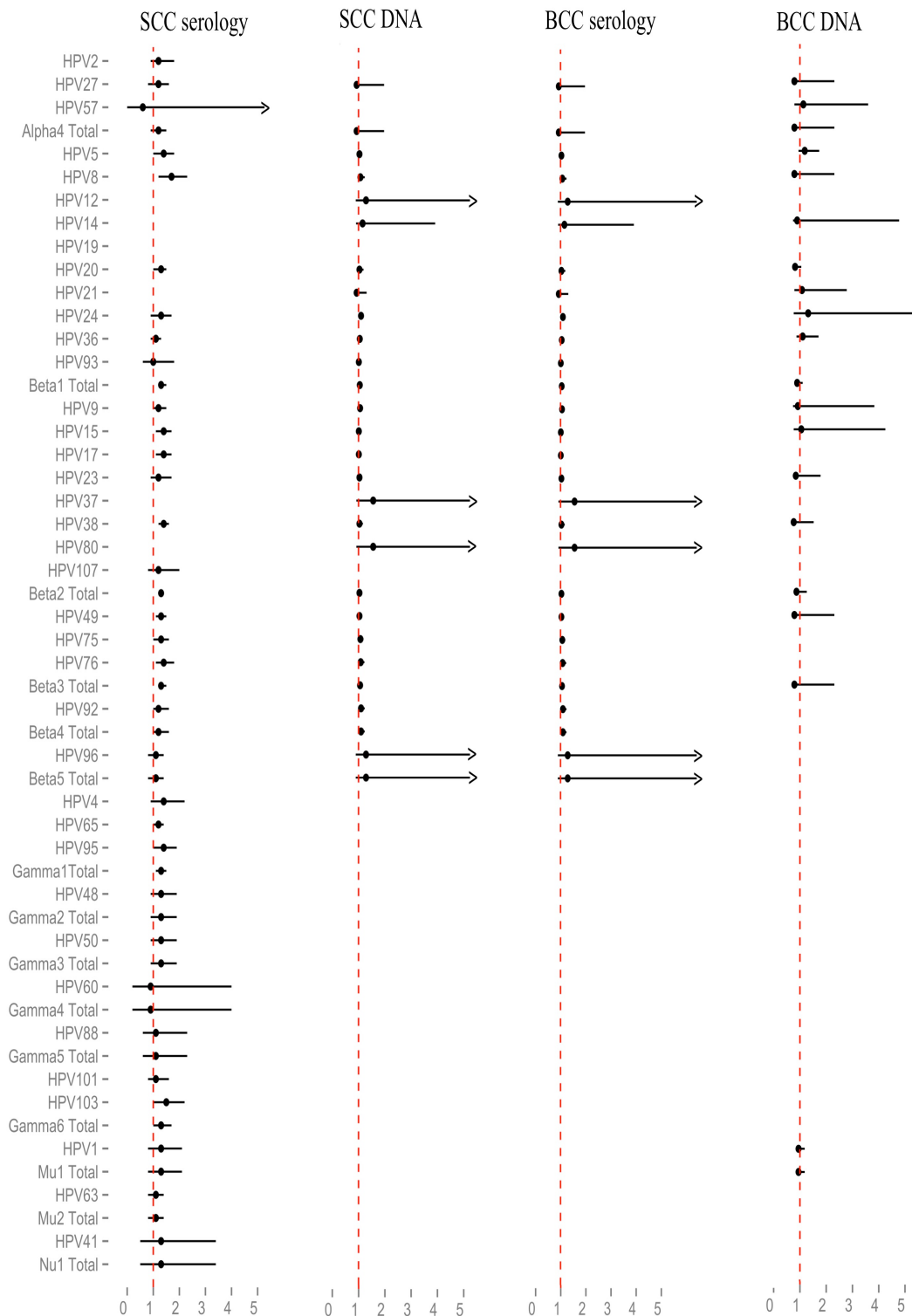


Figure 9. Forest plot of OR estimates of HPV DNA/antibody detectability among SCC/BCC patients and their healthy controls. Black circles represent point estimates of ORs. Black horizontal lines represent 95% CI for respective ORs. If they cross the red dashed line, the result is not statistically significant.

2.3.2 High-throughput NGS technologies in the research on tumor virus epidemiology

2.3.2.1 High-throughput NGS using serum samples

In **Paper II**, serum samples from mothers to leukemic children were subjected to high-throughput NGS. Data was analysed with different bioinformatic pipelines and assembly algorithms. They resulted in a total of 190 non-redundant TTV-related contiguous sequences (table 3). Seventy-eight of them had less than 95% identity with each other. Thus, they were considered as putatively different TTV isolates.

Fifty-eight of the putative isolates were able to produce ≥ 60 aa length *Anelloviridae* proteins after a six frame translation. Forty-six of these protein coding contigs were able to produce an ORF1 protein, while 12 contigs produced only ORF2, ORF3 and ORF4 proteins.

Forty contigs were able to code for an ORF1 protein over $\geq 80\%$ of the length. This was the main criteria to classify them as already known or putatively new types. Twenty-nine of them were classified as putatively new types (figure 10). Analysis of their translation profiles revealed that all of them contained at least one conserved protein signature presumed to be unique for TTVs, which indicates that they truly belong to the *Anelloviridae* family. In **Paper II**, we also identified putatively intragenomic rearranged TTV molecules (figure 10), similar to previously identified cloned TTV isolates from serum samples of mothers to leukemic children [62] (figure 5).

A main concern in NGS studies is that misassembly may construct erroneous “chimeric” contigs that actually belong to different viruses. This could result in findings of putatively rearranged TTV molecules that might be artifacts and actually due to misassembly. 454 GS FLX sequencing technology is susceptible to indel-type errors [162] introducing frame shifts in coding regions. This may have led to the smaller ORFs in putatively rearranged TTV sequences in **Paper II**. Also, for some of the putatively novel TTV types reported in **Paper II**, only subgenomic sequences were obtained. They were not overlapping with each other, thus some of these may represent different parts of the same TTV genome.

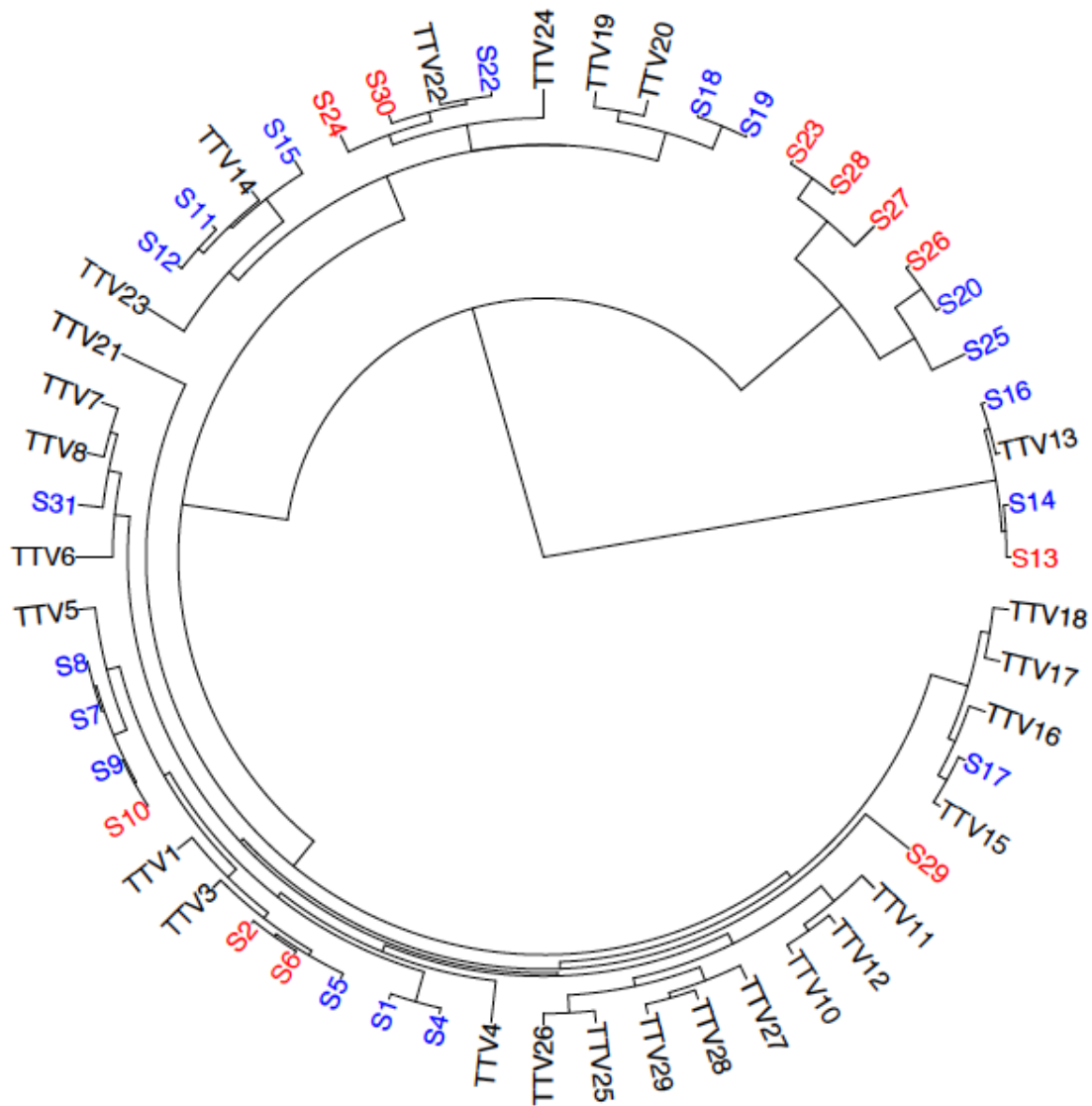


Figure 10. A Bayesian tree based on the 29 putative novel TTV-S sequences (red and blue color) and complete genomes of representative isolates of 29 TTV species groups (black color). Putatively novel TTV-S sequences in red colour represent putatively rearranged TTV molecules.

The results in **Paper II** further extends the knowledge of the extreme genetic diversity of TTVs (figure 10) and provides evidence that the *Anelloviridae* family is much more diverse than what was detectable by conventional molecular detection methods. The large number of different viruses raises the possibility that, even if most TTV infections are harmless, maybe a subset of genotypes in a subset of children might be pathogenic. This hypothesis will be possible to pursue using the methodology of the **Paper II**. TT viruses are probably among the most widespread chronic human viral infections and the most genetically diverse viral group infecting humans. However, there has been relatively little attention and success to develop an efficient methodology to study these viruses.

Table 3. Number of identified contigs stratified by percent identities to closest TT genome of species group (adapted from Paper II).

Species Group	Number of Contigs			Total
	<80%	≥80%<90%	≥90%	
TTV 1	1	0	0	1
TTV 3	20	6	27	53
TTV 5	7	6	0	13
TTV 7	0	3	0	3
TTV 10	0	0	1	1
TTV 15	0	2	3	5
TTV 13	2	8	5	15
TTV 18	0	1	2	3
TTV 19	0	1	0	1
TTV 20	2	0	0	2
TTV 22	9	53	30	92
TTMV 9	1	0	0	1
Total	42	80	68	190

In conclusion, the findings of **Paper II** suggest that high-throughput NGS technology is useful to describe known or unknown anelloviruses that are present in serum samples of pregnant women. Prospective epidemiological studies to investigate possible pathogenicity of these viruses may be warranted.

2.3.2.2 *High-throughput NGS using swab samples, fresh-frozen biopsies and formalin-fixed paraffin-embedded samples*

We performed NGS of amplicons from general primer PCR (FAP primers) for HPVs on samples from putatively HPV associated lesions, such as swab samples from 82 SCCs and 60 AKs, paraffin-embedded biopsies from 28 SCCs and 72 KAs and fresh-frozen biopsies from 92 KAs, 85 SCCs and 92 AKs using GSFLX 454 technology (Roche) (**Paper V**). The NGS revealed an extended diversity of HPV types in these lesions and identified altogether 44 putatively novel HPV types, designated as SE1 to SE44. Later we used bidirectional sequencing using 454 Titanium chemistry and 47 additional putatively novel HPV types were detected [23] in the same samples as in **Paper V** (figure 11).

In **Paper III**, we investigated the presence of virus DNA in a variety of skin lesions (the same samples as in **Paper V**) using NGS but without prior PCR amplification. The amount of DNA in the samples was amplified only using WGA, a method that is independent of any prior knowledge of virus sequences. Unbiased NGS obtained a total of 4284 viral reads (**Paper III**), out of which 4168 were HPV related. Most of them originated from 15 known HPV types (HPV8, HPV12, HPV20, HPV36, HPV38, HPV45, HPV57, HPV59, HPV104, HPV105, HPV107, HPV109, HPV124, HPV138,

HPV147) and four previously described putative types (HPV 915 F 06 007 FD1, FA73, FA101, SE42). **Paper III** also identified two putatively new HPV types SE46 (figure 12) and SE47 (table 4). The putative type SE42 was cloned, sequenced and established as HPV type 155, with only 76% similarity to the most closely related known HPV type. For the putative type FA101, NGS obtained a 7359 bp long contig, representing a complete HPV genome. The complete genome was formed by 247 reads from the pool of paraffin embedded KAs. Non-HPV-related viruses from **Paper III** included human herpesvirus 8, EBV, human endogenous retrovirus, MCV, HPyV6 and TTV.

A similar unbiased approach as in **Paper III** reported a large number of virus related sequence reads, most of them originating from HPV and HPyV when sequencing six samples from healthy forehead skin [12].

We also compared effectiveness of different methods that separated viral DNA from human DNA before WGA in **Paper III**. Directly subjecting samples to WGA and sequencing was most successful for detecting viral DNA (table 1). Possible explanation of this phenomenon could be that handling of low amounts of viral DNA may result in loss of available DNA material.

Table 4. Putatively new HPV types identified by metagenomic sequencing.

Virus Name	Found in sample type	GenBank ID	Genus
SE46	Pool of swabs from SCCs and AKs	JX198657	Gamma
SE47a	Pool of FFPEs from KA	JX198658	Alpha
SE47b	FFPE samples from KA	JX198659	Alpha
SE87 (HPV175)	Swab of condyloma (negative by PCR)	KC108721	Gamma
SE92	Swab of condyloma (negative by PCR)	KC108723	Gamma
SE94	Swab of condyloma (negative by PCR)	KC108724	Gamma
SE95	Swab of condyloma (negative by PCR)	KC108725	Gamma
SE100	Swab of condyloma (negative by PCR)	KC108726	Gamma
SE101	Swab of condyloma (negative by PCR)	KC108727	Gamma
SE102	Swab of condyloma (negative by PCR)	KC108728	Gamma
SE103	Swab of condyloma (negative by PCR)	KC108729	Gamma
SE104	Swab of condyloma (negative by PCR)	KC108730	Gamma
SE105	Swab of condyloma (negative by PCR)	KC108731	Gamma
SE106	Swab of condyloma (negative by PCR)	KC108732	Gamma
SE107	Swab of condyloma (negative by PCR)	KC108733	Gamma
SE109	Swab of condyloma (negative by PCR)	KC108734	Gamma
SE110	Swab of condyloma (negative by PCR)	KC108735	Gamma
SE113	Swab of condyloma (negative by PCR)	KC108736	Gamma
SE114	Swab of condyloma (negative by PCR)	KC108737	Gamma
SE116	Swab of condyloma (negative by PCR)	KC108738	Gamma

In **Paper III**, NGS also detected HPV109 in a pool of skin cancer samples that had previously been negative for HPV by PCR [82]. HPV109 might have been missed by the general primer PCR because this virus has several mismatches in the sequence

corresponding to the “general” primers. In **Paper III**, two novel putative HPV types, SE46 and SE47 (table 4), were detected that had escaped detection when PCR was used prior to NGS in **Paper V**.

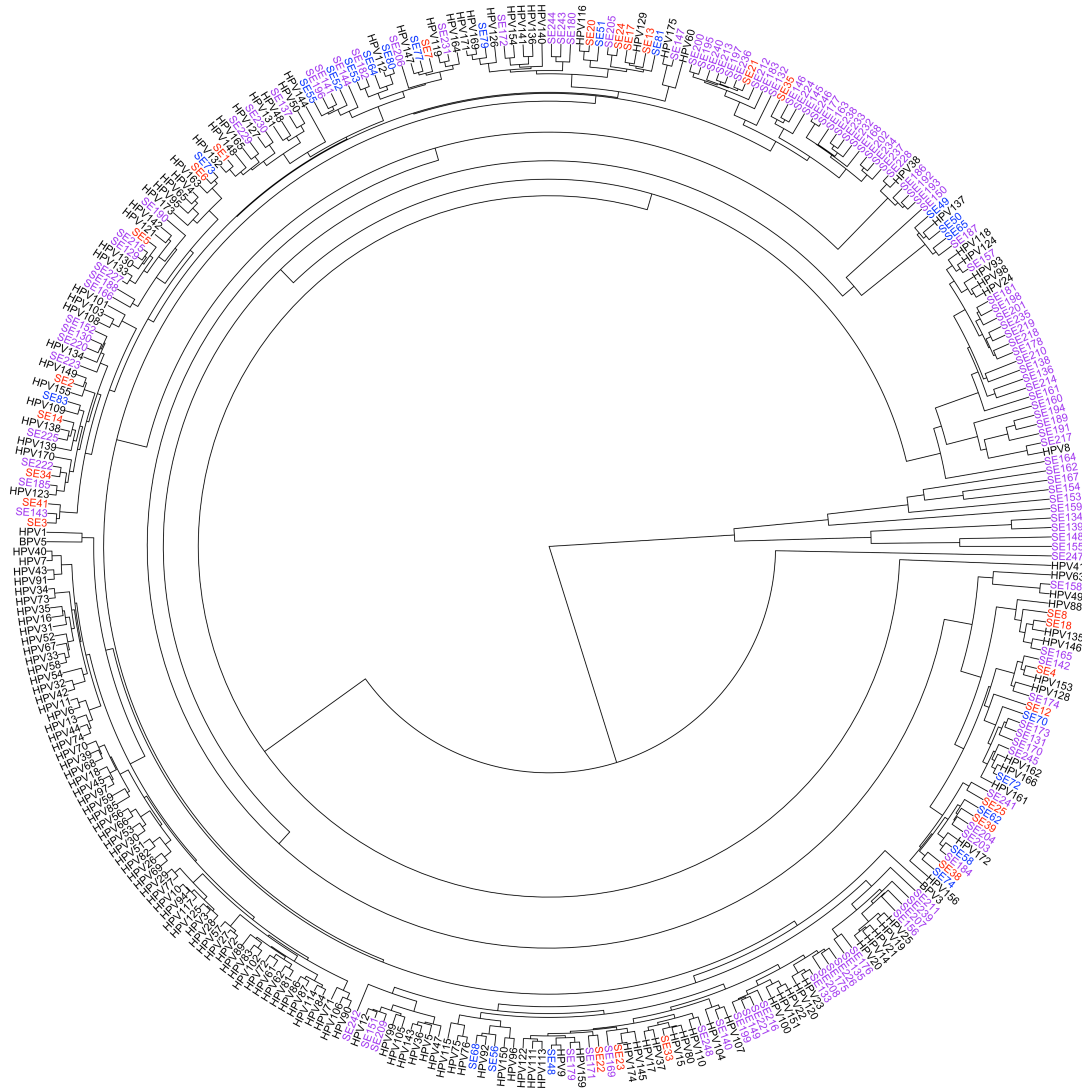


Figure 11. Phylogenetic tree of 164 established HPV types (+ bovine papillomavirus type 3 and type 5) and 160 putative SE types. SE types discovered by GSFLX 454 (Paper V), GSFLX 454 titanium chemistry [23] and Illumina MiSeq (manuscript in preparation) are presented in red, blue, and purple colors, respectively. Phylogenetic tree is based on the L1 part of the complete genomes and 3’ end of putative SE types.

In **Paper VI** we investigated the usefulness of NGS in swab samples from condylomas previously negative for HPVs by conventional PCR methods. Conventional PCR methods may in some case classify condylomas, a disease that is caused by HPV, as “HPV negative”. NGS obtained a total of 4269 reads which had viral origin in such specimens. Among them 1337 (31%) were related to HPVs. Detected HPV-related sequences represented 17 putatively novel gammapapillomaviruses (Table 4), 10 established HPV types (alphapapillomaviruses: HPV6, HPV57, HPV58 and HPV66, betapapillomaviruses: HPV5, HPV105, HPV124, and gammapapillomaviruses HPV50, HPV130, HPV150) and two putative HPV sequences (KC7 and FA69).

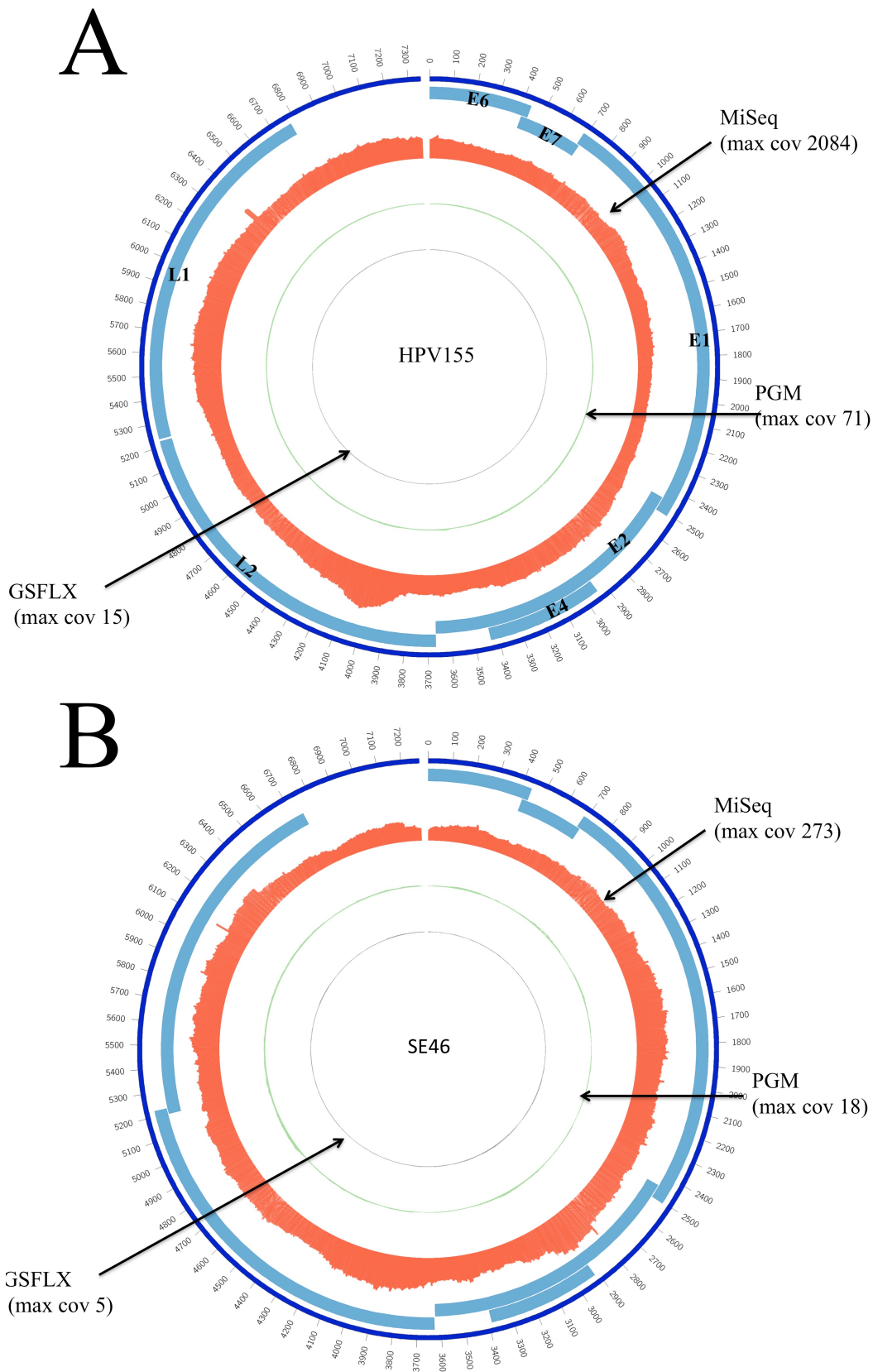


Figure 12. Coverage plots of (A) HPV155 and (B) SE46 genomes from different sequencing runs. Coverage is represented as percentages (maximum coverage/coverage at particular position) comparing different sequencing platforms with each other. Red, green and grey histograms correspond to genome coverages from Illumina MiSeq, Ion Torrent PGM and GSFLX 454, respectively. Light blue lines represent ORFs of HPV155 and putative ORFs of SE46. Plots were generated using Circos visualization tool [27].

Specimens tested in these studies may contain several closely related HPV types, and the possibility exists that assembly algorithms may construct erroneous “chimeric” sequences by the assembly of two different sequences from different HPV viruses. All the HPV sequences reported in **Paper III**, **Paper V**, **Paper VI** were subjected to our “chimera checking” procedure, described in the section on “Bioinformatics for viral metagenomics”, above. Also, some of the reported sub-genomic HPV sequences may represent different parts of the same virus. Indeed seven and six sub genomic SE sequences, reported in **Paper V** and **Paper VI**, respectively, were later identified to belong to the same virus when more complete sequences were obtained [23].

There has been a dramatic evolution of NGS technologies. Figure 12 illustrates two HPVs initially discovered from pools of swab samples from SCC and AK patients by GSFLX 454 sequencing. SE42 (nowadays HPV155) and SE46 were identified with 156 reads (maximum coverage of 15) and 22 reads (maximum coverage of 5) (figure 12), respectively. SE42 (HPV155) was found with a complete genome but due to indel-type errors, it had frame shift errors in the genome. Because of the low coverage we did not manage to reconstruct possible correct ORFs using bioinformatics procedures. SE46 was identified with only a partial 646bp long sequence from the L2 region. Later, the same samples were sequenced on Ion Torrent PGM and Illumina MiSeq. Ion Torrent PGM increased the sequencing depth approximately seven times (HPV155 – 1199 reads; SE46 – 132 reads). SE42 (HPV155) was reconstructed with better quality but a few frame shift errors were still present. However, it was possible to identify error positions and correct them manually after alignment visualisations. SE46 was recovered with longer 950bp and 3774bp long contigs. Illumina MiSeq increased the sequencing depth approximately 220 times for these particular viruses (figure 12). SE42 (HPV155) and SE46 were identified with 33668 (maximum coverage of 2084) and 4881 (maximum coverage of 273) reads, respectively (figure 12). Both SE42 (HPV155) and SE46 were recovered with complete genomes and a genomic organization similar to that of established gammapapillomaviruses. SE42 was later cloned, was named as HPV155 and its genomic organization was confirmed by sequencing with conventional primer walking methods. SE46 is a putative novel type and its genomic organization needs to be confirmed (figure 12). Another clear illustration of NGS development is depicted in figure 11. The HPV general primer FAP amplimers from the same pool of samples were sequenced on GSFLX 454 (**Paper V**), GSFLX 454 with titanium chemistry [23] and Illumina MiSeq platform (manuscript in preparation). GSFLX 454 titanium chemistry [23] doubled the number of identified putatively new HPV types to that of the original GSFLX 454 chemistry (figure 12). Illumina MiSeq platform tripled the identified putatively new HPV types to that of GSFLX454’s original and titanium chemistries together (figure 12).

In conclusion, findings of **Paper V** indicate that the human skin harbours a broad spectrum of different HPV types and that the diversity of HPVs is far greater than we know today. This was confirmed in later investigations when we used improved sequencing chemistry [23] and NGS platform with much larger sequencing depth (Illumina Miseq) (figure 11). Findings of **Paper III** and **Paper VI** suggest that NGS is

a useful technique for unbiased viral DNA detection in swab samples, fresh-frozen biopsies from stripped skin and formalin-fixed paraffin-embedded lesions. Also, it demonstrates an advantage of PCR-free unbiased method, as it will detect the most abundant viruses present without being biased by the PCR primer sequences used (table 4). NGS technology is also rapidly developing and the cost per base is rapidly decreasing.

2.4 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Our findings suggest that patients diagnosed with MCC have an increased risk of second primary cancers compared to the general population (**Study I**). This may be caused by shared etiological factors, such as exposure to UV radiation, MCV infection and/or immunosuppression. Immunosuppression induces an impaired ability to control tumorigenic viruses [3]. Exploration of a possible role of infections in cancer forms that are increased among the immunosuppressed patients but not known to contain viruses is warranted, as pointed out by Harald zur Hausen.

Non-melanoma skin cancers such as SCC are highly increased among immunosuppressed patients [10-13]. The findings of our systematic review (**Study IV**) indicated that HPV species Beta-1, Beta-2, Beta-3 and Gamma-1 had an increased SCC risk. However, on the HPV type level there was a limited number of observations. This finding demonstrates that further research is necessary to clarify the association between the presence of specific HPV types and the risk of SCC.

However, before performing large scale epidemiologic studies it is necessary to investigate, in an unbiased manner, which viruses are present in the tissues from the cases and the controls. Metagenomic sequencing based on NGS was demonstrated to be useful to comprehensively determine which viral DNA is present in different skin lesions (**Study III**) and condylomas (**paper VI**). NGS of HPV general primer FAP amplicons from skin lesions revealed an extended diversity of HPV types in putatively HPV-associated lesions (**Study V**). This indicates that the diversity of HPVs is far greater than what has been detected by conventional methods. The fact that conventional PCR methods are biased towards the primers used was demonstrated in **Paper VI**, when a plethora of HPV types were detected in condylomas, classified as “HPV negative” by PCR.

Even though infectious etiology of childhood leukemias was proposed long time ago [16] there are extremely few studies that have attempted to investigate the association between the two and no specific infectious agent has been identified to date. The most probable explanation for this is the low throughput and bias of conventional molecular detection methods. When serum samples from mothers of leukemic children were subjected to NGS, an extended diversity of anelloviruses was discovered (**Study II**). Thus, it is necessary to detect the complete picture of viral communities present in the biospecimens before a large-scale epidemiological study can be launched. Our findings (**Study II**) suggest that it is feasible to analyse the spectrum of viruses present in human sera by the NGS technology and that the methodology could be generally useful in prospective epidemiological studies of virus-disease associations.

NGS technologies are rapidly evolving, providing increased reliability, increased sequencing depth and lower cost. This, in combination with high quality patient data registries and bio-specimen banks in the Nordic countries offers unique opportunities to

conduct state of the science molecular epidemiological research in the area of infections and cancer.

3 ACKNOWLEDGEMENTS

There are many people who have supported me as a PhD Student and without their support it would have been impossible to accomplish the work presented in this thesis. I am particularly grateful to:

My supervisor **Joakim Dillner**, first of all I want to thank you for trusting me and guiding me into the amazing world of science. Thank you for being an excellent supervisor and friend at the same time. Thank you for letting me stay with your family when I had nowhere to stay. I had many interesting and nice adventures in your research group but I will particularly remember driving a truck from Flensburg to Stockholm. I consider myself extremely lucky for being your student.

My co-supervisor **Bengt Persson**, for guidance and support in the NGS bioinformatics issues, quick response and time for discussions.

Matti Lehtinen, for providing me with enormous support during my first days in the science. As I told you during one of the CCPRB meetings in Tampere you are my godfather in science. Thank you very much for helping me to write my study plan, giving me very important comments on manuscripts and on this thesis.

Ethel-Michele de Villiers, for letting me to work with your research group and helping me to understand TT viruses. Without your help and discussions it would have been impossible to understand anything about these viruses. Thank you very much for your time and for showing me Heidelberg.

Johanna Ekström, Emilie Hultin, Carina Eklund, Sarra Arroyo Muhr, Birgitta Möller for all the excellent data from the lab. Without your help I would had never made it.

Anna-Mari Nykänen, Helena Persson, Anna Olofsson Franzoia, Gudrun Rafi and Helena Anderson for always being kind and helpful in the administrative issues.

Maria Hortlund, for shearing with me your experience about registry-linkages and databases.

All my co-authors, for nice collaborations and for achieving our goals.

Emilie, Carina, Sara, Vitaly and Joakim for wonderful skiing vacations in Idre Fjäll, which was also very nice place to discuss work related issues in relaxed and informal atmosphere.

Carina, for helping me to translate the abstract in to Swedish. **Sara, Emilie and Helena** for going through the thesis for proof checking.

All my colleagues in Malmö lab not mentioned above: **Aline, Anders, Anna, Augustin, Hanna, Kia, Kristina, Kristin, Ola, Olaf**, as well as in Huddinge lab: **Nasrin, Anders, Sara, Camilla and Helena**.

Maria and **Daniel** for the exciting parties in the weekends after long and hard work days.

My Georgian friends **Eka, Kalender** ☺, **Tamara** and **Nino**. While in Malmö lab your apartment **Eka** was my after work destination. So I was almost living in your apartment for 2 years. Now visiting your wonderful family, **Kalender** and **Eka**, is the favourite destination in Malmö for my family.

My brothers **Levani** and **Zura**, for always being there for me.

My **mom** and **dad**, for being helpful and trying to understand me in everything. There is no way I can re-pay you for what you have done for me, but I promise I will try to be the best father for your grand child ☺.

My son **Lukas** for letting me to sleep during nights, especially last few months. As soon as I submit this I will be playing with you.

My wife **Viktoria**, for endless support and taking care of Lukas all your own all the time, especially last month while I was trying to memorise what I was doing last 4 years and write up the thesis.

As I mentioned above there are many people who contributed to this thesis and in case I missed to mention you please don't take it personal. I really appreciate all your support it's just that I am now in a hurry to submit the thesis for printing.

4 REFERENCES

1. Rous P (1911) A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. *J Exp Med* 13: 397-411.
2. Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, et al. (2009) A review of human carcinogens--Part B: biological agents. *Lancet Oncol* 10: 321-322.
3. (2012) Biological agents. Volume 100 B. A review of human carcinogens. IARC Monogr Eval Carcinog Risks Hum 100: 1-441.
4. Feng H, Shuda M, Chang Y, Moore PS (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319: 1096-1100.
5. Shuda M, Feng H, Kwun HJ, Rosen ST, Gjoerup O, et al. (2008) T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. *Proc Natl Acad Sci U S A* 105: 16272-16277.
6. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, et al. (2012) Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol* 13: 607-615.
7. de Martel C, Franceschi S (2009) Infections and cancer: established associations and new hypotheses. *Crit Rev Oncol Hematol* 70: 183-194.
8. Hemminki K, Dillner J (2009) Editorial. *Int J Cancer* 125: vii.
9. Schulz TF (2009) Cancer and viral infections in immunocompromised individuals. *Int J Cancer* 125: 1755-1763.
10. Boukamp P (2005) Non-melanoma skin cancer: what drives tumor development and progression? *Carcinogenesis* 26: 1657-1667.
11. Grulich AE, van Leeuwen MT, Falster MO, Vajdic CM (2007) Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. *Lancet* 370: 59-67.
12. Lindelof B, Sigurgeirsson B, Gabel H, Stern RS (2000) Incidence of skin cancer in 5356 patients following organ transplantation. *Br J Dermatol* 143: 513-519.
13. Berg D, Otley CC (2002) Skin cancer in organ transplant recipients: Epidemiology, pathogenesis, and management. *J Am Acad Dermatol* 47: 1-17; quiz 18-20.
14. Moloney FJ, Comber H, O'Lorcain P, O'Kelly P, Conlon PJ, et al. (2006) A population-based study of skin cancer incidence and prevalence in renal transplant recipients. *Br J Dermatol* 154: 498-504.
15. Hartevelt MM, Bavinck JN, Kootte AM, Vermeer BJ, Vandenbroucke JP (1990) Incidence of skin cancer after renal transplantation in The Netherlands. *Transplantation* 49: 506-509.
16. Kinlen LJ (1998) Infection and childhood leukemia. *Cancer Causes Control* 9: 237-239.
17. Laara E (2011) Study designs for biobank-based epidemiologic research on chronic diseases. *Methods Mol Biol* 675: 165-178.
18. Pukkala E (2011) Nordic biological specimen bank cohorts as basis for studies of cancer causes and control: quality control tools for study cohorts with more than two million sample donors and 130,000 prospective cancers. *Methods Mol Biol* 675: 61-112.
19. Bzhalava D, Johansson H, Ekstrom J, Faust H, Moller B, et al. (2013) Unbiased Approach For Virus Detection In Skin Lesions. *Plos One* 8: e65953.
20. Johansson H, Bzhalava D, Ekstrom J, Hultin E, Dillner J, et al. (2013) Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. *Virology*.
21. Foulongne V, Sauvage V, Hebert C, Dereure O, Cheval J, et al. (2012) Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS One* 7: e38499.
22. Ekstrom J, Bzhalava D, Svenback D, Forslund O, Dillner J (2011) High throughput sequencing reveals diversity of Human Papillomaviruses in cutaneous lesions. *Int J Cancer* 129: 2643-2650.

23. Ekstrom J, Muhr LS, Bzhalava D, Soderlund-Strand A, Hultin E, et al. (2013) Diversity of human papillomaviruses in skin lesions. *Virology* 447: 300-311.
24. de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H (2004) Classification of papillomaviruses. *Virology* 324: 17-27.
25. Bernard HU, Burk RD, Chen Z, van Doorslaer K, Hausen H, et al. (2010) Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* 401: 70-79.
26. (2007) Human papillomaviruses. *IARC Monogr Eval Carcinog Risks Hum* 90: 1-636.
27. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19: 1639-1645.
28. Duensing S, Munger K (2004) Mechanisms of genomic instability in human cancer: insights from studies with human papillomavirus oncoproteins. *Int J Cancer* 109: 157-162.
29. Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, et al. (2003) Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med* 348: 518-527.
30. Plummer M, Schiffman M, Castle PE, Maucort-Boulch D, Wheeler CM (2007) A 2-year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. *J Infect Dis* 195: 1582-1589.
31. Vaccarella S, Herrero R, Dai M, Snijders PJ, Meijer CJ, et al. (2006) Reproductive factors, oral contraceptive use, and human papillomavirus infection: pooled analysis of the IARC HPV prevalence surveys. *Cancer Epidemiol Biomarkers Prev* 15: 2148-2153.
32. Vaccarella S, Herrero R, Snijders PJ, Dai M, Thomas JO, et al. (2008) Smoking and human papillomavirus infection: pooled analysis of the International Agency for Research on Cancer HPV Prevalence Surveys. *Int J Epidemiol* 37: 536-546.
33. Hernandez BY, Wilkens LR, Zhu X, McDuffie K, Thompson P, et al. (2008) Circumcision and human papillomavirus infection in men: a site-specific comparison. *J Infect Dis* 197: 787-794.
34. Clifford GM, Goncalves MA, Franceschi S (2006) Human papillomavirus types among women infected with HIV: a meta-analysis. *AIDS* 20: 2337-2344.
35. Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, et al. (1999) Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol* 189: 12-19.
36. Smith JS, Lindsay L, Hoots B, Keys J, Franceschi S, et al. (2007) Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. *Int J Cancer* 121: 621-632.
37. Curado MP, Edwards B, Shin HR, Ferlay J, Heanue M, et al. (2008) Cancer incidence in five continents. Volume IX. *IARC Sci Publ*: 1-837.
38. Marur S, D'Souza G, Westra WH, Forastiere AA (2010) HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol* 11: 781-789.
39. Schiffman M, Clifford G, Buonaguro FM (2009) Classification of weakly carcinogenic human papillomavirus types: addressing the limits of epidemiology at the borderline. *Infect Agent Cancer* 4: 8.
40. Wallace NA, Robinson K, Howie HL, Galloway DA (2012) HPV 5 and 8 E6 abrogate ATR activity resulting in increased persistence of UVB induced DNA damage. *PLoS Pathog* 8: e1002807.
41. Viarisis D, Mueller-Decker K, Kloz U, Aengeneyndt B, Kopp-Schneider A, et al. (2011) E6 and E7 from beta HPV38 cooperate with ultraviolet light in the development of actinic keratosis-like lesions and squamous cell carcinoma in mice. *PLoS Pathog* 7: e1002125.
42. Accardi R, Dong W, Smet A, Cui R, Hautefeuille A, et al. (2006) Skin human papillomavirus type 38 alters p53 functions by accumulation of deltaNp73. *EMBO Rep* 7: 334-340.
43. Saidj D, Cros MP, Hernandez-Vargas H, Guarino F, Sylla BS, et al. (2013) Oncoprotein E7 from beta human papillomavirus 38 induces formation of an

- inhibitory complex for a subset of p53-regulated promoters. *J Virol* 87: 12139-12150.
44. Howie HL, Koop JI, Weese J, Robinson K, Wipf G, et al. (2011) Beta-HPV 5 and 8 E6 promote p300 degradation by blocking AKT/p300 association. *PLoS Pathog* 7: e1002211.
 45. Muench P, Probst S, Schuetz J, Leiprecht N, Busch M, et al. (2010) Cutaneous papillomavirus E6 proteins must interact with p300 and block p53-mediated apoptosis for cellular immortalization and tumorigenesis. *Cancer Res* 70: 6913-6924.
 46. Asgari MM, Kiviat NB, Crichtlow CW, Stern JE, Argenyi ZB, et al. (2008) Detection of human papillomavirus DNA in cutaneous squamous cell carcinoma among immunocompetent individuals. *J Invest Dermatol* 128: 1409-1417.
 47. Patel AS, Karagas MR, Perry AE, Nelson HH (2008) Exposure profiles and human papillomavirus infection in skin cancer: an analysis of 25 genus beta-types in a population-based study. *J Invest Dermatol* 128: 2888-2893.
 48. Plasmeyer EI, Neale RE, de Koning MN, Quint WG, McBride P, et al. (2009) Persistence of betapapillomavirus infections as a risk factor for actinic keratoses, precursor to cutaneous squamous cell carcinoma. *Cancer Res* 69: 8926-8931.
 49. Andersson K, Michael KM, Luostarinen T, Waterboer T, Gislefoss R, et al. (2012) Prospective study of human papillomavirus seropositivity and risk of nonmelanoma skin cancer. *Am J Epidemiol* 175: 685-695.
 50. Waterboer T, Abeni D, Sampogna F, Rother A, Masini C, et al. (2008) Serological association of beta and gamma human papillomaviruses with squamous cell carcinoma of the skin. *Br J Dermatol* 159: 457-459.
 51. Bzhalava D, Guan P, Franceschi S, Dillner J, Clifford G (2013) A systematic review of the prevalence of mucosal and cutaneous human papillomavirus types. *Virology* 445: 224-231.
 52. Okamoto H (2009) History of discoveries and pathogenicity of TT viruses. *Curr Top Microbiol Immunol* 331: 1-20.
 53. Biagini P (2009) Classification of TTV and related viruses (anelloviruses). *Curr Top Microbiol Immunol* 331: 21-33.
 54. Jelcic I, Hotz-Wagenblatt A, Hunziker A, Zur Hausen H, de Villiers EM (2004) Isolation of multiple TT virus genotypes from spleen biopsy tissue from a Hodgkin's disease patient: genome reorganization and diversity in the hypervariable region. *J Virol* 78: 7498-7507.
 55. Mushahwar IK, Erker JC, Muerhoff AS, Leary TP, Simons JN, et al. (1999) Molecular and biophysical characterization of TT virus: evidence for a new virus family infecting humans. *Proc Natl Acad Sci U S A* 96: 3177-3182.
 56. Niagro FD, Forsthoefel AN, Lawther RP, Kamalanathan L, Ritchie BW, et al. (1998) Beak and feather disease virus and porcine circovirus genomes: intermediates between the geminiviruses and plant circoviruses. *Arch Virol* 143: 1723-1744.
 57. Hijikata M, Takahashi K, Mishiro S (1999) Complete circular DNA genome of a TT virus variant (isolate name SANBAN) and 44 partial ORF2 sequences implicating a great degree of diversity beyond genotypes. *Virology* 260: 17-22.
 58. Noteborn MH, Koch G (1995) Chicken anaemia virus infection: molecular basis of pathogenicity. *Avian Pathol* 24: 11-31.
 59. Okamoto H, Nishizawa T, Tawara A, Peng Y, Takahashi M, et al. (2000) Species-specific TT viruses in humans and nonhuman primates and their phylogenetic relatedness. *Virology* 277: 368-378.
 60. Okamoto H, Takahashi M, Nishizawa T, Tawara A, Fukai K, et al. (2002) Genomic characterization of TT viruses (TTVs) in pigs, cats and dogs and their relatedness with species-specific TTVs in primates and tupaia. *J Gen Virol* 83: 1291-1297.
 61. de Villiers EM, Kimmel R, Leppik L, Gunst K (2009) Intragenomic rearrangement in TT viruses: a possible role in the pathogenesis of disease. *Curr Top Microbiol Immunol* 331: 91-107.

62. Leppik L, Gunst K, Lehtinen M, Dillner J, Streker K, et al. (2007) In vivo and in vitro intragenomic rearrangement of TT viruses. *J Virol* 81: 9346-9356.
63. Garbuglia AR, Iezzi T, Capobianchi MR, Pignoloni P, Pulsoni A, et al. (2003) Detection of TT virus in lymph node biopsies of B-cell lymphoma and Hodgkin's disease, and its association with EBV infection. *Int J Immunopathol Pharmacol* 16: 109-118.
64. Shiramizu B, Yu Q, Hu N, Yanagihara R, Nerurkar VR (2002) Investigation of TT virus in the etiology of pediatric acute lymphoblastic leukemia. *Pediatr Hematol Oncol* 19: 543-551.
65. Pineau P, Meddeb M, Raselli R, Qin LX, Terris B, et al. (2000) Effect of TT virus infection on hepatocellular carcinoma development: results of a Euro-Asian survey. *J Infect Dis* 181: 1138-1142.
66. Parkin DM, Bray F (2006) Chapter 2: The burden of HPV-related cancers. *Vaccine* 24 Suppl 3: S3/11-25.
67. Stanley M (2003) Chapter 17: Genital human papillomavirus infections--current and prospective therapies. *J Natl Cancer Inst Monogr*: 117-124.
68. Baseman JG, Koutsky LA (2005) The epidemiology of human papillomavirus infections. *J Clin Virol* 32 Suppl 1: S16-24.
69. Bosch FX, Lorincz A, Munoz N, Meijer CJ, Shah KV (2002) The causal relation between human papillomavirus and cervical cancer. *J Clin Pathol* 55: 244-265.
70. Clifford GM, Smith JS, Plummer M, Munoz N, Franceschi S (2003) Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. *Br J Cancer* 88: 63-73.
71. Kapeu AS, Luostarinen T, Jellum E, Dillner J, Hakama M, et al. (2009) Is smoking an independent risk factor for invasive cervical cancer? A nested case-control study within Nordic biobanks. *Am J Epidemiol* 169: 480-488.
72. (2006) Cervical carcinoma and reproductive factors: collaborative reanalysis of individual data on 16,563 women with cervical carcinoma and 33,542 women without cervical carcinoma from 25 epidemiological studies. *Int J Cancer* 119: 1108-1124.
73. (2009) Cervical carcinoma and sexual behavior: collaborative reanalysis of individual data on 15,461 women with cervical carcinoma and 29,164 women without cervical carcinoma from 21 epidemiological studies. *Cancer Epidemiol Biomarkers Prev* 18: 1060-1069.
74. Smith JS, Herrero R, Bosetti C, Munoz N, Bosch FX, et al. (2002) Herpes simplex virus-2 as a human papillomavirus cofactor in the etiology of invasive cervical cancer. *J Natl Cancer Inst* 94: 1604-1613.
75. Koskela P, Anttila T, Bjorge T, Brunsvig A, Dillner J, et al. (2000) Chlamydia trachomatis infection as a risk factor for invasive cervical cancer. *Int J Cancer* 85: 35-39.
76. Castle PE, Giuliano AR (2003) Chapter 4: Genital tract infections, cervical inflammation, and antioxidant nutrients--assessing their roles as human papillomavirus cofactors. *J Natl Cancer Inst Monogr*: 29-34.
77. Oberyshyn TM (2008) Non-melanoma skin cancer: importance of gender, immunosuppressive status and vitamin D. *Cancer Lett* 261: 127-136.
78. (2012) Cancer Incidence in Sweden 2011.
79. Liao PB (2008) Merkel cell carcinoma. *Dermatol Ther* 21: 447-451.
80. Hussain SK, Sundquist J, Hemminki K (2010) Incidence trends of squamous cell and rare skin cancers in the Swedish national cancer registry point to calendar year and age-dependent increases. *J Invest Dermatol* 130: 1323-1328.
81. Zwald FO, Brown M (2011) Skin cancer in solid organ transplant recipients: advances in therapy and management: part I. Epidemiology of skin cancer in solid organ transplant recipients. *J Am Acad Dermatol* 65: 253-261; quiz 262.
82. Forslund O, Iftner T, Andersson K, Lindelof B, Hradil E, et al. (2007) Cutaneous human papillomaviruses found in sun-exposed skin: Beta-papillomavirus species 2 predominates in squamous cell carcinoma. *J Infect Dis* 196: 876-883.
83. Jablonska S, Majewski S (1994) Epidermodysplasia verruciformis: immunological and clinical aspects. *Curr Top Microbiol Immunol* 186: 157-175.
84. Pfister H, Ter Schegget J (1997) Role of HPV in cutaneous premalignant and malignant tumors. *Clin Dermatol* 15: 335-347.

85. Forslund O, DeAngelis PM, Beigi M, Schjolberg AR, Clausen OP (2003) Identification of human papillomavirus in keratoacanthomas. *J Cutan Pathol* 30: 423-429.
86. Farzan SF, Waterboer T, Gui J, Nelson HH, Li Z, et al. (2013) Cutaneous alpha, beta and gamma human papillomaviruses in relation to squamous cell carcinoma of the skin: a population-based study. *Int J Cancer* 133: 1713-1720.
87. Iannacone MR, Gheit T, Waterboer T, Giuliano AR, Messina JL, et al. (2012) Case-control study of cutaneous human papillomaviruses in squamous cell carcinoma of the skin. *Cancer Epidemiol Biomarkers Prev* 21: 1303-1313.
88. Karagas MR, Waterboer T, Li Z, Nelson HH, Michael KM, et al. (2010) Genus beta human papillomaviruses and incidence of basal cell and squamous cell carcinomas of skin: population based case-control study. *BMJ* 341: c2986.
89. Bouwes Bavinck JN, Neale RE, Abeni D, Euvrard S, Green AC, et al. (2010) Multicenter study of the association between betapapillomavirus infection and cutaneous squamous cell carcinoma. *Cancer Res* 70: 9777-9786.
90. Struijk L, Hall L, van der Meijden E, Wanningen P, Bavinck JN, et al. (2006) Markers of cutaneous human papillomavirus infection in individuals with tumor-free skin, actinic keratoses, and squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev* 15: 529-535.
91. Vasiljevic N, Hazard K, Dillner J, Forslund O (2008) Four novel human betapapillomaviruses of species 2 preferentially found in actinic keratosis. *J Gen Virol* 89: 2467-2474.
92. Weissenborn SJ, Nindl I, Purdie K, Harwood C, Proby C, et al. (2005) Human papillomavirus-DNA loads in actinic keratoses exceed those in non-melanoma skin cancers. *J Invest Dermatol* 125: 93-97.
93. Antonsson A, Erfurt C, Hazard K, Holmgren V, Simon M, et al. (2003) Prevalence and type spectrum of human papillomaviruses in healthy skin samples collected in three continents. *J Gen Virol* 84: 1881-1886.
94. de Koning MN, Weissenborn SJ, Abeni D, Bouwes Bavinck JN, Euvrard S, et al. (2009) Prevalence and associated factors of betapapillomavirus infections in individuals without cutaneous squamous cell carcinoma. *J Gen Virol* 90: 1611-1621.
95. Hazard K, Karlsson A, Andersson K, Ekberg H, Dillner J, et al. (2007) Cutaneous human papillomaviruses persist on healthy skin. *J Invest Dermatol* 127: 116-119.
96. Alam M, Ratner D (2001) Cutaneous squamous-cell carcinoma. *N Engl J Med* 344: 975-983.
97. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB (2010) Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host Microbe* 7: 509-515.
98. Moens U, Ludvigsen M, Van Ghelue M (2011) Human polyomaviruses in skin diseases. *Patholog Res Int* 2011: 123491.
99. Loyo M, Guerrero-Preston R, Brait M, Hoque MO, Chuang A, et al. (2010) Quantitative detection of Merkel cell virus in human tissues and possible mode of transmission. *Int J Cancer* 126: 2991-2996.
100. Shuda M, Arora R, Kwun HJ, Feng H, Sarid R, et al. (2009) Human Merkel cell polyomavirus infection I. MCV T antigen expression in Merkel cell carcinoma, lymphoid tissues and lymphoid tumors. *Int J Cancer* 125: 1243-1249.
101. Carter JJ, Paulson KG, Wipf GC, Miranda D, Madeleine MM, et al. (2009) Association of Merkel cell polyomavirus-specific antibodies with Merkel cell carcinoma. *J Natl Cancer Inst* 101: 1510-1522.
102. Pastrana DV, Tolstov YL, Becker JC, Moore PS, Chang Y, et al. (2009) Quantitation of human seroresponsiveness to Merkel cell polyomavirus. *PLoS Pathog* 5: e1000578.
103. Tolstov YL, Pastrana DV, Feng H, Becker JC, Jenkins FJ, et al. (2009) Human Merkel cell polyomavirus infection II. MCV is a common human infection that can be detected by conformational capsid epitope immunoassays. *Int J Cancer* 125: 1250-1256.

104. Wieland U, Mauch C, Kreuter A, Krieg T, Pfister H (2009) Merkel cell polyomavirus DNA in persons without merkel cell carcinoma. *Emerg Infect Dis* 15: 1496-1498.
105. Faust H, Andersson K, Ekstrom J, Hortlund M, Robsahm TE, et al. (2014) Prospective study of Merkel cell polyomavirus and risk of Merkel cell carcinoma. *Int J Cancer* 134: 844-848.
106. Look AT (1997) Oncogenic transcription factors in the human acute leukemias. *Science* 278: 1059-1064.
107. Greaves MF, Wiemels J (2003) Origins of chromosome translocations in childhood leukaemia. *Nat Rev Cancer* 3: 639-649.
108. zur Hausen H, de Villiers EM (2005) Virus target cell conditioning model to explain some epidemiologic characteristics of childhood leukemias and lymphomas. *Int J Cancer* 115: 1-5.
109. Petridou E, Dalamaga M, Mentis A, Skalkidou A, Moustaki M, et al. (2001) Evidence on the infectious etiology of childhood leukemia: the role of low herd immunity (Greece). *Cancer Causes Control* 12: 645-652.
110. Hakulinen T, Hovi L, Karkinen J, Penttinen K, Saxen L (1973) Association between influenza during pregnancy and childhood leukaemia. *Br Med J* 4: 265-267.
111. Curnen MG, Varma AA, Christine BW, Turgeon LR (1974) Childhood leukemia and maternal infectious diseases during pregnancy. *J Natl Cancer Inst* 53: 943-947.
112. Austin DF, Karp S, Dworsky R, Henderson BE (1975) Excess leukemia in cohorts of children born following influenza epidemics. *Am J Epidemiol* 101: 77-83.
113. Vianna NJ, Polan AK (1976) Childhood lymphatic leukemia: prenatal seasonality and possible association with congenital varicella. *Am J Epidemiol* 103: 321-332.
114. Lehtinen M, Koskela P, Ogmundsdottir HM, Bloigu A, Dillner J, et al. (2003) Maternal herpesvirus infections and risk of acute lymphoblastic leukemia in the offspring. *Am J Epidemiol* 158: 207-213.
115. Lehtinen M, Ogmundsdottir HM, Bloigu A, Hakulinen T, Hemminki E, et al. (2005) Associations between three types of maternal bacterial infection and risk of leukemia in the offspring. *Am J Epidemiol* 162: 662-667.
116. Tedeschi R, Bloigu A, Ogmundsdottir HM, Marus A, Dillner J, et al. (2007) Activation of maternal Epstein-Barr virus infection and risk of acute leukemia in the offspring. *Am J Epidemiol* 165: 134-137.
117. Tedeschi R, Luostarinen T, Marus A, Bzhalava D, Ogmundsdottir HM, et al. (2009) No risk of maternal EBV infection for childhood leukemia. *Cancer Epidemiol Biomarkers Prev* 18: 2790-2792.
118. Fedrick J, Alberman ED (1972) Reported influenza in pregnancy and subsequent cancer in the child. *Br Med J* 2: 485-488.
119. Dockerty JD, Skegg DC, Elwood JM, Herbison GP, Becroft DM, et al. (1999) Infections, vaccinations, and the risk of childhood leukaemia. *Br J Cancer* 80: 1483-1489.
120. Kwan ML, Metayer C, Crouse V, Buffler PA (2007) Maternal illness and drug/medication use during the period surrounding pregnancy and risk of childhood leukemia among offspring. *Am J Epidemiol* 165: 27-35.
121. van Steensel-Moll HA, Valkenburg HA, Vandenbroucke JP, van Zanen GE (1985) Are maternal fertility problems related to childhood leukaemia? *Int J Epidemiol* 14: 555-559.
122. Roman E, Ansell P, Bull D (1997) Leukaemia and non-Hodgkin's lymphoma in children and young adults: are prenatal and neonatal factors important determinants of disease? *Br J Cancer* 76: 406-415.
123. McKinney PA, Juszczak E, Findlay E, Smith K, Thomson CS (1999) Pre- and perinatal risk factors for childhood leukaemia and other malignancies: a Scottish case control study. *Br J Cancer* 80: 1844-1851.
124. Naumburg E, Bellocco R, Cnattingius S, Jonzon A, Ekblom A (2002) Perinatal exposure to infection and risk of childhood leukemia. *Med Pediatr Oncol* 38: 391-397.

125. Gustafsson B, Huang W, Bogdanovic G, Gauffin F, Nordgren A, et al. (2007) Adenovirus DNA is detected at increased frequency in Guthrie cards from children who develop acute lymphoblastic leukaemia. *Br J Cancer* 97: 992-994.
126. Honkaniemi E, Talekar G, Huang W, Bogdanovic G, Forestier E, et al. (2010) Adenovirus DNA in Guthrie cards from children who develop acute lymphoblastic leukaemia (ALL). *Br J Cancer* 102: 796-798.
127. Petridou E, Revinthi K, Alexander FE, Haidas S, Kolioukas D, et al. (1996) Space-time clustering of childhood leukaemia in Greece: evidence supporting a viral aetiology. *Br J Cancer* 73: 1278-1283.
128. Pukkala E, Andersen A, Berglund G, Gislefoss R, Gudnason V, et al. (2007) Nordic biological specimen banks as basis for studies of cancer causes and control--more than 2 million sample donors, 25 million person years and 100,000 prospective cancers. *Acta Oncol* 46: 286-307.
129. Tulinius H, Storm HH, Pukkala E, Andersen A, Ericsson J (1992) Cancer in the Nordic countries, 1981-86. A joint publication of the five Nordic Cancer Registries. *APMIS Suppl* 31: 1-194.
130. Bzhalava D, Ekstrom J, Lysholm F, Hultin E, Faust H, et al. (2012) Phylogenetically diverse TT virus viremia among pregnant women. *Virology* 432: 427-434.
131. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207-214.
132. (2012) A framework for human microbiome research. *Nature* 486: 215-221.
133. Wylie KM, Weinstock GM, Storch GA (2012) Emerging view of the human virome. *Transl Res* 160: 283-290.
134. Lecuit M, Eloit M (2013) The human virome: new tools and concepts. *Trends Microbiol* 21: 510-515.
135. Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA (2012) Sequence analysis of the human virome in febrile and afebrile children. *PLoS One* 7: e27735.
136. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One* 4: e7370.
137. Lazarczyk M, Cassonnet P, Pons C, Jacob Y, Favre M (2009) The EVER proteins as a natural barrier against papillomaviruses: a new insight into the pathogenesis of human papillomavirus infections. *Microbiol Mol Biol Rev* 73: 348-370.
138. Bhattarai N, Stapleton JT (2012) GB virus C: the good boy virus? *Trends Microbiol* 20: 124-130.
139. Meiring TL, Salimo AT, Coetzee B, Maree HJ, Moodley J, et al. (2012) Next-generation sequencing of cervical DNA detects human papillomavirus types not detected by commercial kits. *Virol J* 9: 164.
140. Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* 2: 3.
141. Towner JS, Sealy TK, Khristova ML, Albarino CG, Conlan S, et al. (2008) Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog* 4: e1000212.
142. Willner D, Haynes MR, Furlan M, Hanson N, Kirby B, et al. (2012) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. *Am J Respir Cell Mol Biol* 46: 127-131.
143. Johansson H, Bzhalava D, Ekstrom J, Hultin E, Dillner J, et al. (2013) Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. *Virology* 440: 1-7.
144. Moore PS, Chang Y (2010) Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat Rev Cancer* 10: 878-889.
145. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186-194.
146. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, et al. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 10: 57-59.

147. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 11: 187.
148. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3: 1314-1317.
149. Hutchison CA, 3rd, Smith HO, Pfannkoch C, Venter JC (2005) Cell-free cloning using phi29 DNA polymerase. *Proc Natl Acad Sci U S A* 102: 17332-17336.
150. Fancello L, Raoult D, Desnues C (2012) Computational tools for viral metagenomics and their application in clinical research. *Virology* 434: 162-174.
151. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17: 377-386.
152. Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, et al. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* 5: e1000593.
153. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One* 6: e27992.
154. Lindner MS, Renard BY (2013) Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res* 41: e10.
155. Deeks JJ, Higgins JP, Altman DG (2008) Analysing Data and Undertaking Meta-Analyses, in *Cochrane Handbook for Systematic Reviews of Interventions*; Higgins JP, Green S, editors. Chichester, UK: John Wiley & Sons, Ltd.
156. Greenland S, O' Rourke K (2008) Meta-Analysis. In: Rothman KJ, Greenland S, Lash T, editors. *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott Williams and Wilkins. pp. 652-682.
157. Thompson SG (1994) Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 309: 1351-1355.
158. Thompson SG, Higgins JP (2002) How should meta-regression analyses be undertaken and interpreted? *Stat Med* 21: 1559-1573.
159. Howard RA, Dores GM, Curtis RE, Anderson WF, Travis LB (2006) Merkel cell carcinoma and multiple primary cancers. *Cancer Epidemiol Biomarkers Prev* 15: 1545-1549.
160. Koljonen V, Kukko H, Tukiainen E, Bohling T, Sankila R, et al. (2010) Second cancers following the diagnosis of Merkel cell carcinoma: a nationwide cohort study. *Cancer Epidemiol* 34: 62-65.
161. Schiffman M, Herrero R, Desalle R, Hildesheim A, Wacholder S, et al. (2005) The carcinogenicity of human papillomavirus types reflects viral evolution. *Virology* 337: 76-84.
162. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.