Analysis of genetic susceptibility to cervical cancer using candidate gene and GWAS approaches

IVANA JUKO-PECIREP
Abstract


Cervical cancer is the forth most commonly diagnosed cancer among women worldwide. It is caused by persistent infection with an oncogenic type of Human Papillomavirus (HPV). The HPV is a necessary but not sufficient cause of cervical cancer. Environmental factors such as smoking, high parity and long-term use of oral contraceptives increases the risk of cervical cancer. Genetic factors also affect the risk of developing the disease. The aim of this thesis is to search for and evaluate genetic risk factors for cervical cancer using both a candidate gene approach and a genome-wide association study (GWAS).

Paper I examined the association of genetic variation in three Fanconi Anemia (FA) genes (FANCA, FANCC and FANCL), involved in DNA repair, with cervical cancer susceptibility in the Swedish population. No association was observed. Paper II evaluated the association of genetic variation in the TMC6 and TMC8 genes with susceptibility to cervical cancer in the Swedish population and an association of two SNPs (rs2290907 and rs16970849) with cervical cancer was observed. In paper III the first GWAS performed in cervical cancer was reported. Three independent loci in the major histocompatibility complex (MHC) region at 6p21.3 were found to affect the susceptibility to cervical cancer. Paper IV examined the sequence variation in the TMC6 and TMC8 region and its association with cervical cancer. A highly polymorphic 21 bp sequence was identified and found to be repeated 5 to 42 times in both cases and controls. Lack of this repeat was associated with increased risk of cervical cancer. An intronic SNP (rs2926778) located in between the TNRC6C and TMC6 genes was also found to be associated with cervical cancer.

The thesis provides evidence for the importance of genes in the immune system for cervical cancer susceptibility. The genetic risk factors identified explain only a part of the genetic susceptibility, implying that other risk factors remains to be identified.

Keywords: cervical cancer, association study, human papillomavirus, genetics, complex disease, TMC6, TMC8, MHC region

Ivana Juko-Pecirep, Department of Immunology, Genetics and Pathology, Rudbecklaboratoriet, Uppsala University, SE-751 85 Uppsala, Sweden.

© Ivana Juko-Pecirep 2015

ISSN 1651-6206
urn:nbn:se:uu:diva-248484 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-248484)
Till min kär familj ♥ Mojog dragoj obitelji
"Promise me you'll always remember: You're braver than you believe, and stronger than you seem, and smarter than you think"
Winnie The Pooh
List of Papers

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


Reprints were made with permission from the respective publisher.
Related articles

# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td>Cervical cancer</td>
<td>13</td>
</tr>
<tr>
<td>Screening and treatment of cervical cancer</td>
<td>15</td>
</tr>
<tr>
<td>Prevention of cervical cancer</td>
<td>17</td>
</tr>
<tr>
<td>Human Papillomavirus</td>
<td>18</td>
</tr>
<tr>
<td>Additional risk factors</td>
<td>20</td>
</tr>
<tr>
<td>Environmental risk factors</td>
<td>20</td>
</tr>
<tr>
<td>Genetic risk factors</td>
<td>20</td>
</tr>
<tr>
<td>Genetic analysis of complex diseases</td>
<td>23</td>
</tr>
<tr>
<td>Cohort vs. Case-Control study</td>
<td>23</td>
</tr>
<tr>
<td>Familial aggregation</td>
<td>23</td>
</tr>
<tr>
<td>Segregation study</td>
<td>23</td>
</tr>
<tr>
<td>Linkage study</td>
<td>23</td>
</tr>
<tr>
<td>Association study</td>
<td>24</td>
</tr>
<tr>
<td>Technologies in complex disease analysis</td>
<td>25</td>
</tr>
<tr>
<td>Genotyping methods</td>
<td>25</td>
</tr>
<tr>
<td>Sanger Sequencing</td>
<td>27</td>
</tr>
<tr>
<td>Next Generation Sequencing</td>
<td>27</td>
</tr>
<tr>
<td>Present investigation</td>
<td>30</td>
</tr>
<tr>
<td>Aim</td>
<td>30</td>
</tr>
<tr>
<td>Paper I</td>
<td>30</td>
</tr>
<tr>
<td>Background</td>
<td>30</td>
</tr>
<tr>
<td>Results and discussions</td>
<td>30</td>
</tr>
<tr>
<td>Paper II</td>
<td>31</td>
</tr>
<tr>
<td>Background</td>
<td>31</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>31</td>
</tr>
<tr>
<td>Paper III</td>
<td>32</td>
</tr>
<tr>
<td>Background</td>
<td>32</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>32</td>
</tr>
<tr>
<td>Paper IV</td>
<td>34</td>
</tr>
<tr>
<td>Background</td>
<td>34</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>35</td>
</tr>
<tr>
<td>Conclusion and future perspectives</td>
<td>37</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>39</td>
</tr>
<tr>
<td>References</td>
<td>41</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ASC</td>
<td>Atypical squamous cells</td>
</tr>
<tr>
<td>ASO</td>
<td>Allele-specific oligos</td>
</tr>
<tr>
<td>CIS</td>
<td>Cancer In Situ</td>
</tr>
<tr>
<td>CIN</td>
<td>Cervical intraepithelial neoplasm</td>
</tr>
<tr>
<td>CMI</td>
<td>Cell-mediated immunity</td>
</tr>
<tr>
<td>CE</td>
<td>Capillary electrophoresis</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxynucleotide triphosphates</td>
</tr>
<tr>
<td>EV</td>
<td>Epidermodysplasia Verruciformis</td>
</tr>
<tr>
<td>HR</td>
<td>High-risk</td>
</tr>
<tr>
<td>TS</td>
<td>Tumor suppressor</td>
</tr>
<tr>
<td>HPV</td>
<td>Human Papillomavirus</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>SCC</td>
<td>Squamous cell carcinoma</td>
</tr>
<tr>
<td>Pap</td>
<td>Papanicolaou-stain</td>
</tr>
<tr>
<td>LSIL</td>
<td>Low-grade squamous intraepithelial lesions</td>
</tr>
<tr>
<td>LR</td>
<td>Low-risk</td>
</tr>
<tr>
<td>HSIL</td>
<td>High-grade squamous intraepithelial lesions</td>
</tr>
<tr>
<td>LEEP</td>
<td>Loop electrosurgical excision procedure</td>
</tr>
<tr>
<td>RCT</td>
<td>Randomized control trial</td>
</tr>
<tr>
<td>URR</td>
<td>Upstream regulatory region</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>MHC</td>
<td>Major histocompatibility complex</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigen</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td>FANC</td>
<td>Fanconi anemia complementation group</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>LSO</td>
<td>Locus-specific oligo</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>MPS</td>
<td>Massively parallel sequencing</td>
</tr>
<tr>
<td>SMRT</td>
<td>Single-molecule real-time</td>
</tr>
</tbody>
</table>
Introduction

Cancer is caused by uncontrolled growth of cells. Abnormal proliferation is called neoplasia and results in neoplasm, also called tumor. Tumors may be benign or malignant; where the major difference is that the malignant tumors destroy adjacent normal tissue and sometimes spreads to other parts of the body (metastases). Cancers are also classified by type of cells that the tumor starts in and is therefore presumed to be the origin of the tumor. The most common cancers are those of epithelial origin and are classified as carcinomas. Development of cancer progress is divided in three steps. Dysplasia is the earliest form of pre-cancerous lesion and can be classified in low-grade or high-grade. Low-grade dysplasia means that some atypical cell changes are seen, but not in all cells, while in high-grade dysplasia atypical changes are seen in many cells as well as abnormal growth. The second step is called carcinoma in situ (CIS), a Latin word for ”in its place” meaning that the surrounding tissue has not been invaded. CIS is synonymous with high-grade dysplasia and the risk of transforming into invasive cancer growth is high. The final step is invasive carcinoma, commonly called cancer that will, if left untreated, spread beyond the tissue in which it developed and grow into surrounding, healthy tissue and may be lethal.

Development of cancer is a multistep process, where normal somatic cells acquires different abilities in a stepwise fashion and evolves into a cancer cell. A single mutation is generally not enough to convert a normal cell into a malignant one, but a number of independent events are necessary in order for malignant transformation to take place. The cells need to be self-sufficient in growth factors and insensitive to signals that normally inhibit growth. They also need to avoid apoptosis and overcome the replication limit that normal cells have. Finally the developing tumor needs the ability to invade the surrounding tissue and metastasize. For all these changes the transforming cells need to overcome the natural defense mechanisms against cancer.

Cancer is a genetic disease, meaning that they result from unnatural function of one or more genes. It should not be mixed up with the statement that cancer is a hereditary disease. A hereditary disease is passed from the parents to a child through the inheritance of a defective gene. However, some rare cancers are hereditary, for most of them the genes could be an affecting component for the cancer development and for some of them it just is a by chance event.
The alterations described above require genetic changes in key genes. There are three categories of genes that are targets for these mutations: oncogenes, tumor suppressor genes and DNA repair genes. The normal function of oncogenes is to control cell proliferation, apoptosis, or both. The product of oncogenes can be divided into six broad groups: transcription factors, chromatin remodelers, growth factors, growth factor receptors, signal transducers and apoptosis regulators. The oncogenes can be activated by chromosomal rearrangements, point mutations or gene amplification, causing alteration in oncogene structure or an increase/decrease of its expression. The genetic changes results in a growth advantage or increased survival of cells carrying such alteration. Tumor suppressor (TS) genes protect the cell from events leading to cancer. In a normally functioning cell some TS gene products prevent inappropriate cell cycle progression, some steer deviant cells into apoptosis, while other keep the genome stable and mutation rates low by ensuring accurate replication, repair and segregation of the cells DNA. Unlike oncogenes, both alleles of a tumor suppressor gene must be inactivated in order to cause a change in the behavior of the cell. DNA repair proteins are often classified as TS genes, mainly because mutations in these genes increase the risk of cancer development.

The work presented in this thesis is focusing on evaluating a common cancer in women, namely cervical cancer caused by the human papilloma-virus (HPV). Using candidate gene and genome wide association studies (GWAS) approaches we are evaluating the genetic influence on the disease.
Cervical cancer

The cervix is found in the lower part of the uterus; the cervical canal connects the uterus and the vagina. (Figure 1) The epithelium of the cervix varies, consisting of two types of epithelial cells, squamous cells that form layers in the epithelium and columnar cells that form the glandular epithelium. The junction where these two cell types meet is called the transformation zone. In this zone the columnar epithelium transforms into squamous epithelium and this is the site where dysplasia, the first step of cervical carcinogenesis, develops. Cervical cancer is divided into two types: Cervical squamous cell carcinoma, which is derived from squamous cells and cervical adenocarcinoma, arising in the glandular cells of cervix.

*Figure 1. A schematic drawing of the female reproductive organs. The arrow is indicating the location of the cervix, found between uterus and vagina. Illustration by Erik Gustavsson*
528 000 new cancer cases and 266 000 deaths due to cervical cancer was registered. Around 85% of new cancer cases occurred in the less developed countries, where also nine out of ten (87%) cervical cancer deaths occurred. Globally, the highest incidence rates of cervical cancer are reported in Eastern Africa, Melanesia, Southern and Middle Africa and the lowest rates in Australia/New Zealand and Western Asia. The average age for developing cervical cancer is 55 years, however one-fourth of the women diagnosed with cervical cancer are younger than 40 years. The cervical squamous cell carcinoma (SCC) is the most common form, and for a couple of years ago accounted for 75% of all cervical neoplasms in developing countries. Adenocarcinomas and adenosquamous cell carcinomas represent 10-25%, where other rare types comprise the remaining cases.

The reasons for high incidence rates in developing countries may be due to different prevalence of risk factors, which will be discussed later in the thesis. However, the major reason is believed to be the lack of cervical cancer screening and diagnostic methods, which is an important step in detection of precancerous and early stage cervical cancer.

*Figure 2.* The incidence of cervical cancer cases (pink) and deaths (red) in Sweden, has decreased with 50% during the 1960-2012.
Screening and treatment of cervical cancer

In Sweden organized cervical cancer screening was implemented in the mid-1960s, even before the cause of cervical cancer was known, and has reduced the numbers of cervical cancer cases with 50%. (Figure 2) According to the Swedish guidelines all women between ages 23-60 are invited to cytological screening every 3rd year. Women after 50 between 60 are invited with a 5 years interval. In the screening program cytological differences of cells in transformation zone of cervix are examined using Papanicolaou-stained (Pap) smear, that was introduced 1949. In Sweden 680 000 Pap smears are taken every year. Of them are more than 30 000 women diagnosed with cytological abnormalities. In 2008, 3 450 Swedish women were diagnosed with cancer in situ of the cervix. In 2007, 466 women were diagnosed with cervical cancer, giving an incidence rate of 10.1 cases per 100 000.

The cells from cervix are collected with a cytobrush, spread out on a microscope slide and stained to enable examination by the microscope. The Pap-smear is graded depending on the appearance of the cells. The cervical intraepithelial neoplasm (CIN) system was introduced 1973 and is a scale to grade cytological differences. Cell abnormalities are graded from one to three, processing from mild cervical intraepithelial neoplasia (CIN1) to more moderate or severe degree of neoplasia and micro invasive lesions (CIN2 or CIN3). In the early 1990 the Bethesda system was developed to reflect an advanced understanding of cervical neoplasia and introduce uniform descriptive diagnostic histological terminology. The system was modified in 2001 and today the cytological squamous cell differences are classified into four categories: (1) ASC (atypical squamous cells), (2) LSIL (low-grade squamous intraepithelial lesions) (3) HSIL (high-grade squamous intraepithelial lesions), and (4) squamous cell carcinoma.

However, even if Pap smear is used as primary method for detection of high-risk HPV induced cell changes, the method has its limitations. False negatives rates as high as 20-30% have been reported and 8% of the samples received are inadequate. If the Pap smear shows abnormal findings but the patients do not have gross cervical lesions, they are often evaluated by colposcopy and colposcopy-directed biopsy. Colposcopy is an optical examination of the mucous membrane in the cervix and helps to establish where the cell change has occurred and how progressive it is. A biopsy is taken from the area suspected to be the origin for an examination with microscope. In the biopsy the presence of HPV DNA or RNA can be shown by in situ hybridization with probes labeled with either radioisotopes or chemically reactive ligands, which are detected by autoradiography, fluorescence or a detection of color reaction. The biopsy is as well examined for dysplasia. If classified as CIN2 or higher (CIN2+) conization is performed, where the mucosal membrane together with the surrounding tissue in the transformation zone is removed with a carbon dioxide (CO2) laser, a surgical knife or loop electro-
surgical excision procedure (LEEP). In 95% out of all cases one of the mentioned procedures is necessary to remove and cure cell changes\textsuperscript{15}.

So far the primary diagnostic tools have been cytology and histology, but new molecular methods detecting HPV DNA in clinical samples are rapidly replacing cytology as the primary test in screening. The US Food and Drug Administration (FDA) in 2014 approved HPV DNA to be used for testing of women age 25 and older. If the woman is positive for HPV 16 or HPV 18 a colposcopy exam is conducted. Women positive for any of the other 12 high risk HPVs should have a Pap test to determine the need for a colposcopy\textsuperscript{16}. Previously only women who’s Pap smear was classified as atypical squamous of unknown significance (ASCUS) were reexamined with an HPV DNA test\textsuperscript{17}. The importance of HPV testing as support to the screening program has been evaluated in several studies\textsuperscript{18, 19}. The National Board of Health and Welfare in Sweden are suggesting changes in the Swedish screening systems where HPV DNA test is recommended as the primary screening test from 2017. In women age 23-29 a cytology test every third year should still be considered, because there is no proof that HPV DNA test would be of a higher effectiveness. This age group has an overall higher prevalence of HPV infections. That could lead to an unnecessary evaluation due to that the infection is often self-healing. However, women age 39-49 should be invited every third year for a HPV DNA test because of higher incidence of cell changes and cervical cancer. In the age group 50-64 the examination should as well be with an HPV DNA test but every seventh year. This since the prevalence of cell changes and cervical cancer are not that common and that the HPV DNA tests have longer duration than the cytological test\textsuperscript{20}.

Which of the different tools for diagnosis of cervical cancer is to prefer based on the knowledge so far? Studies comparing the Pap-smear and HPV DNA testing indicate that HPV DNA testing has better sensitivity and negative predictive values, as well as specificity\textsuperscript{18, 19, 21, 22}. At present the exact HPV types are not reported clinically, they are mainly used in large epidemiology and research studies. Still, HPV testing can be of high value, due to that not only the presence of HPV but also the viral load is noted, which has been proposed as a marker for persistent infection and risk of development of cervical lesions\textsuperscript{23}.
Prevention of cervical cancer

Screening for cell changes and prevention of HPV infection are both important strategies in cervical cancer prevention. The possibility of treating the disease at an early stage often disrupts the process of neoplasia.

Preventing the HPV infection can prevent cervical cancer. Today this can be achieved by usage of the HPV prophylactic vaccines. Gardasil® and Cervarix®, both consisting of the major viral coat protein L1, assembled into macromolecular structures are very similar to the wild-type viron. Still the vaccines are not identical. Gardasil® that was introduces in 2006 is quadrivalent, protecting the patient against HPV-6, -11, -16 and -18, targeting females 9-26 years of age. The vaccine has also the potential of preventing genital warts. In 2008 a second vaccine was introduced, named Cervarix® and is a bivalent vaccine protecting against HPV-16 and -18. It is recommended for use for females aged 10-25 years and is believed to protect against anogenital warts caused by HPV, precancerous lesions and cervical cancer. In a large phase III trial conducted for four years with randomized placebo-controls, both vaccines were shown to be well tolerated and effective in preventing the disease caused by each vaccine HPV type. In natural infection and vaccination, an immune response and memory of that response protects against infection and disease on re-encounter with the pathogen. It is known that it may take 25-30 years from that a woman is infected with a carcinogenic HPV type until invasive cervical cancer develops. The first pre-invasive cervical lesions are detectible approximately five to seven years after a carcinogenic HPV infection. However, cohorts have been followed for around 10 years for the quadrivalent vaccine and 8.4 years for the bivalent vaccine. Where the data suggest that protection is likely to be long lasting. However, the issue still remains that the long-term protection is unknown; therefore long-term follow-up studies are necessary.

Today, these two HPV vaccines are approved in more than 100 countries. The Swedish National Board of Health and Welfare (NBH) decided to include the HPV vaccine in January 2010 in the childhood vaccination targeting girls at age 10-12. The introduction of the HPV vaccine was approved only if an extensive follow-up program was implemented. The World Health Organization (WHO) has also recommendation of follow-up for coverage, safety and population effectiveness and the aim is to ensure that the screening program continues to be at least as effective as today.

Scientific evidence and the results of the large phase III randomized control trial (RCT) indicate that if the HPV vaccines are introduced to girls and women before the first expose to the HPV virus, it will prevent disease caused by the HPV types included in the vaccine. Still it is very central to remember that even if a woman is vaccinated it is important to attend the screening program. The vaccine protects from the HPV types included, but
there are several other high-risk HPV types that can cause cancer. Another issue is the length of immune memory, up to ten years is known, but what happens after that? However, if we would like to monitor the population to see if the prevalence of the HPV types included in the vaccine decrease, it is necessary to use type specific HPV tests in the screening program.

Human Papillomavirus

In 1976 professor Harald zur Hausen suggested the human papillomavirus to be the causative agent of cervical cancer\(^{35}\). This was also identified in the beginning of the 1980s in human genital warts and cervical biopsies\(^{36}\). In 2008 he was awarded with the Nobel Price in Medicine for his discovery. Papillomaviruses are highly diverse, belonging to the family Papillomaviridae and occur in most mammals and birds. To date over 100 types have been detected in humans and described based on isolation of the complete genomes. About 40 HPV types can infect the epithelial and mucosa lining of the anogenital tract and other areas\(^{37}\). The genital and mucosal types belong to the taxonomy genera alpha-HPV that also contains some cutaneous types, which cause warts\(^{38}\).

Molecular epidemiologic studies have shown that some types of HPV’s, around 18 of them\(^{39}\), are the principal cause of invasive cervical cancer and cervical intraepithelial neoplasia\(^{40}\). Genital HPV types are divided in low-risk (LR), which are found mainly in genital warts, and high-risk (HR) types, which are associated with invasive cervical cancer\(^{41}\). In 2005 IARC (WHO International Agency for research on Cancer) concluded that HPV types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 and 82 are carcinogenic in humans\(^{42}\). Of them HPV 16 and 18 are responsible for around 90% of all cervical cancer cases\(^{40}\). It is also known that infections with more than one HR-HPV type can occur and represents a possible reason for development of the cervical cancer\(^{43}\). Case–control studies, case series, and prevalence surveys have shown that HPV DNA can be detected in 90–100% of all cases. However, there have been some HPV negative cervical cancers reported. These finding are believed to be due to an artifact in the current detection methods, or potentially due to loss of HPV DNA during the progression to cervical cancer\(^{44}\). In contrast, the prevalence of HPV DNA in women identified as suitable epidemiological controls is between 5–20% in cervical cell samples\(^{45}\). Based on this infection with at least one oncogenic high-risk HPV type is regarded as a necessary but not sufficient cause of cervical cancer. HPV is a common sexually transmitted infection and every second women will during here lifetime get exposed and infected with HPV\(^{46}\). The prevalence in sexually active young women ranges from 20-46%\(^{47}\) and decreases with age, but the HPV prevalence is still high in older women and the major
risk factor for cervical cancer development\textsuperscript{48-50}. However, the majority (70-90\%) of women infected will clear the HPV within 12 to 30 months\textsuperscript{51-53}.

HPV is a circular double-stranded DNA virus with a size of 8 kb\textsuperscript{54}. (Figure 3) The HPV genome consists of 3 general regions; an upstream regulatory region (URR) containing sequences that control viral transcription and replication, an early region containing open reading frames (ORFs; e.g. E1, E2, E4, E5, E6 and E7) encoding proteins that are involved in multiple functions like trans-activation of transcription, transformation, replication, and viral adaption to different cellular environments, and a third region called late region that is coding for the L1 and L2 capsid proteins that form the structure of the viron and make the viral DNA packaging and maturation possible\textsuperscript{38,55}.

![Figure 3. A schematic drawing showing the Human Papillomavirus (HPV) genome.](image)

For HPV to successfully infect the host, it requires presence of epidermal and mucosal epithelial cells that are still able to proliferate\textsuperscript{56}. In the basal cell layers the virus only express the early genes (E5, E6 and E7), leading to enhanced proliferation of infected cells and their lateral expansion. The next step is to infect the suprabasal layer, were the virus express the late genes, initiating replication of the circular viral genome and formation of the structural proteins. As the virus reaches the upper layers of the epidermis, or mucosa, complete viral particles are assembled and released. In 1989, Munger et al. showed that only the E6 and E7 genes of HR-HPV types were able to immortalize human cells in tissue culture\textsuperscript{57}. The E6 and E7 genes are consistently expressed in malignant tissue, and inhibiting their expression blocks the malignant phenotype of cervical cancer cells\textsuperscript{58}. This is partly achieved by
the E6 interaction with p53\textsuperscript{59} and E7 with RB\textsuperscript{60}, leading to inactivation of these tumor suppressor pathways.

Additional risk factors

What is a risk factor? A risk factor is anything that changes your chance of getting a disease, in this case cervical cancer. However, having one or several risk factors does not have to mean that you will develop the disease. HPV genital infection is common and HR-HPV is known to be the major risk factor for cervical cancer development. Approximately 70-90\% of all infected women will clear the infection and only a small fraction of HPV infected women will develop cervical cancer\textsuperscript{46}. Therefore other risk factors beside the HPV infection must also be involved in determining the outcome of an HPV infection.

Environmental risk factors

Over the past years, a number of additional environmental risk factors have been evaluated and associated with the risk of transition from a cervical HPV infection to cervical malignancy. However, some of these are assumed to only reflect the risk of acquiring HPV infection such as number of sexual partners because the HPV is spread through genital contact\textsuperscript{61, 62}. Other risk factors associated with HPV infection is an early sexual debut (<18 years), earlier age at first full-time pregnancy (<18 years), high parity\textsuperscript{63}, use of combined hormonal oral contraceptives for longer than 5-years\textsuperscript{64} as well as previous infections with sexually transmitted diseases. Such as human immunodeficiency virus (HIV), herpes simplex virus 2 and Chlamydia trachomatis\textsuperscript{61,65}. Use of tobacco, increases the risk of squamous cell cervical cancer\textsuperscript{66}. The risk rises with quantity of cigarettes smoked per day and number of years smoked\textsuperscript{61}. The explanation could be that smokers are less able to clear the HPV infection or that smoking causes cancerous progression in HPV infected cells\textsuperscript{66}.

Genetic risk factors

Another reason why some women develop cervical cancer, while other clear the HPV infection, could be their genetic constitution. In 1999, Magnusson \textit{et al.} reported a significant familial aggregation for cervical cancer. When comparing incidence of dysplasia and CIS in relatives of women with disease and in controls, they saw that biological relatives have a doubled risk of developing cervical cancers compared to non-relatives. This provides strong epidemiological evidence for a genetic link to the development of cervical cancer\textsuperscript{67}. Later on Magnusson \textit{et al.} studied adoptive and full mothers, as
well as full and adoptive sisters of cases with cervical cancers and estimated the relative importance of shared genes to 27% and shared familial environments to 2%\textsuperscript{68}. This demonstrates that development of cervical cancer depends to a significant degree on inherited genetic factors. Several other studies have also pointed out the importance of genetic risk factors in cervical cancer, where first and second-degree relatives and full and half-siblings were studied. Estimates shows that the familial risk has a heritable component of 71% and an environmental component of 29% in first and second degree relative, and a 64% heritable component and 36% environmental component in full and half-siblings\textsuperscript{69, 70}. As mentioned there is large variation in the susceptibility to developing cervical cancer among women exposed to HR-HPV. This could reflect a difference in host susceptibility to either exposure of HPV, infection following exposure, persistence of the infection or sensitivity to viral oncoproteins\textsuperscript{71}.

Genes of the immune response system have been studied to investigate the potential association with cervical cancer as well as effect on the susceptibility toward HPV infection and the persistence of the infection. Studies have been focused on the major histocompatibility complex (MHC) region; particularly on the human leukocyte antigen (HLA) class II that are a part of the MHC region\textsuperscript{72} and several associations have been identified\textsuperscript{73-76}. The genes found in this region are very polymorphic giving rise to large variety of expressed MHC molecules. The function of the MHC is to display antigen peptides on the cell surface to initiate the cell-mediated immunity (CMI). HLA class I (HLA-A, B and C) are expressed by nucleated cells and present intracellular antigens to CD8+ T cells where HLA class II (HLA.DR, DQ and DP) are expressed by dendritic cells, B cells and macrophages and present phagocytosed extracellular antigens to CD4+ T cells. (Figure 4) So in conclusion a protein can only induce an immune reaction if the MHC molecules present the peptides. If the MHC molecules were absent it would lead to lack of CMI activation.

\textit{Figure 4.} Schematic overview of the MHC region on chromosome 6.
However, several other candidate genes that are part of different pathways have been suggested to influence cervical cancer. Genes found in the MHC regions except of the HLA such as TNF\textsuperscript{77}, LTA\textsuperscript{78}, TAP1 and TAP2\textsuperscript{79}. There are as well other genes in the immunological pathways that have been proposed for instance IL-10\textsuperscript{80}, CCR-2\textsuperscript{81}, IFN\textsubscript{γ}\textsuperscript{82}, KIR\textsuperscript{83}, CD83\textsuperscript{84}, Fas and Fas ligand\textsuperscript{85}. In 2009 Wang \textit{et al.} investigated 92 polymorphisms in 49 immune response and DNA repair genes where they suggested association of the DNA repair gene Fanconi anemia complementation group A (FANCA) with cervical cancer susceptibility.\textsuperscript{86} Following year in 2010 they investigated several different pathways based on their hypothesized function such as DNA repair, viral infection and cell entry. Among the different associations the two genes transmembrane channel-like protein (TMC) 6 and 8 were of particular interest due to their previous association with Epidermodysplasia Verruciformis (EV). That is a disease characterized by increased sensitivity to cutaneous HPV infections and high risk of squamous cell carcinoma of the skin\textsuperscript{87}. In summary, many studies are found suggesting the immune response involvement in the cervical cancer development as well as non-immune response genes. Could it be that it is a combination of genes affecting the outcome of an HPV infection rather than single loci? (Figure 5)

\textit{Figure 5.} A schematic drawing of the risk factors for cervical cancer development. (From Emma Ivansson)
Genetic analysis of complex diseases

Epidemiology studies causes, distribution and control of disease in populations as well as the relationship between causes and effects. Genetic epidemiology focuses on studies of genetic predisposition to disease and the joint effects of genetic and non-genetic (environmental) effects on disease risk. This type of research aims to identify links between disease and the factors that increase the risk of disease. Depending on what kind of scientific question you want to answer, different study approaches are used. The most common study designs are outlined below.

Cohort vs. Case-Control study

Two basic types of study designs in genetic epidemiology are cohort and case–control studies. Cohort study examines one or more health effects of exposure to a single agent. Initially healthy subjects are followed over time to determine the incidence of health outcomes. Classical case–control study examines a single disease in relation to exposure to one or more agents. Here the cases are subjects’ affected by the disease of interest, and the controls are healthy subjects from the investigated population. The purpose of the control group is to provide information on the exposure distribution in the population that gave rise to the cases. Researchers obtain and compare exposure histories of cases as well as controls.

Familial aggregation

Familial aggregation studies evaluate traits, behaviors or disorders within one family that could be caused by genetic or environmental factors.

Segregation study

Segregation studies are performed to evaluate whether the pattern of disease manifestation in families fits to a particular type of inheritance, usually if the inheritance is dominant or recessive.

Linkage study

In genetic epidemiology, analysis of linkage and association are the two most common means of testing for genetic loci influencing risk of disease. Linkage analysis identifies the location of genes that affect the risk of developing a genetic disease. Information on the genetic recombination events in family-based pedigree structures is the basis of linkage studies. This approach has been shown to be very powerful for simple Mendelian diseases.
with rare, severe, and high-risk mutations. The disadvantages of linkage analyses are that variants with fairly small or moderate effects are not likely to be detected. Also, if multiple genes are involved in different families, linkage analyses may not be able to detect them. In complex diseases both multiple genes and environmental, each with small or moderate effects, as well as interaction among them, will impact risk of disease and linkage analysis will not be enough to detect these. In this case association studies may be a better choice.

Association study

Association studies are comparable to case-control studies, the difference is that the disease associated “exposures” that the researcher is interested in to identify are different genetic alleles. Association studies aims to evaluate the correlation between genetic variants and a disease phenotype in a population. Instead of assessing the recombination events, the linkage disequilibrium (LD) is investigated. LD describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from random formation of haplotypes based on the alleles frequencies. If an association is observed, a particular allele, genotype or haplotype will be found more often, or less often, than expected by chance among individuals carrying the trait. Basically the frequencies of variant alleles among affected individuals are compared to unaffected individuals.

There are two possible study designs in association studies. The commonly used one is a case-control study were the distribution of genotype or allele frequencies at a locus in affected subjects is compared with that in unaffected controls, and the statistical analysis is testing if there is significance difference between the two studied groups. If a difference in the frequency of a genotype or allele is observed between the two groups studied, it indicates that the locus is involved in disease susceptibility, or could be in LD with a causal polymorphism. A population-based control study, which is based on sampling cases and unrelated controls, has the advantage that it is easy to collect sufficiently large sample sizes and it therefore has a high statistical power to detect small or moderate effects. However, the drawback with this study design is that the genotype and haplotype frequencies may vary between ethnic or geographic populations. To avoid this, the cases and controls most therefore be matched for ethnicity or geographic origin, as well as other important factors, or else false positive associations can occur because of the confounding effects of population stratification.

The second approach is family-based association studies, were parents or unaffected siblings are used as controls for the case, which is their affected offspring/sibling. This kind of study is used to avoid the potential confounding effects of population stratification, mentioned above, giving the ad
vantage that all individuals in a family pedigree have a common genetic background. However, avoiding one problem leads to another; family-based studies are often small because of the difficulty to collect large samples of well-characterized families and this often leads to lower statistical power. Also, family-based studies are prone to the consequences of a common shared environment. However, comparing the power of the population-based and the family-based approach in a meta-analysis study, no difference was observed between them.

The work presented in this thesis is based on two very common types of association studies. The first one is the candidate gene study approach, which focuses on investigating associations between genetic variation within pre-specified genes of interest and phenotypes or disease states. Candidate genes are selected based on prior knowledge of the genes biological functional impact on the trait or disease question, which is the major difficulty with this approach, or because they have been implicated in disease in previous studies. The advantage of candidate gene studies is that they are better suited for detection of genetic effects for common and more complex diseases where the risk associated with any given candidate gene is relatively small. The second type of design is the genome-wide association study (GWAS), where the entire genome is searched for genes that are association to the disease. Here 100 000 to million of common genetic variants are investigated in large set of patients and controls to search for variants associated with a trait. GWAS most commonly focus on finding associations between single nucleotide polymorphisms (SNPs) and disease state. The scanning of the genome can be done using various study designs, such as case-control studies, cohort studies and clinical trials. The drawback of GWAS is that large samples sizes are necessary for the scanning, due to the very high number of statistical tests of associations (minimum one per SNP) that are performed, requiring a very stringent threshold of statistical significance. Associations discovered by GWAS raise additional questions, particularly because observed effects are typically very small, and the SNPs found to be associated represents markers and further investigation are required to identify the true causal variants.

Technologies in complex disease analysis

Genotyping methods

In a candidate gene study, the SNPs studied for the association with a particular disease can be genotyped using a variety of different typing methods. The SNPs could be genotyped using a Taqman assay, based on the use of real-time polymerase chain reaction (PCR) and the 5’nuclease assay. The
technique is used to amplify and simultaneously distinguish the two alleles at the target gene. The assay includes PCR primers that will amplify the targeted site and a set of two probes that are complementary to the targeted sequence but differs at the site where the SNP is found. The probes are allele-specific and are labeled with different reporter dyes. In addition, both probes are labeled with a quencher dye that will block (quench) the fluorescence from the allele-specific probe. Upon hybridization of the allele-specific oligonucleotide to its complementary sequence, the oligonucleotide is degraded when the polymerase extends the primer sequence during the PCR, fluorophore is released and the florescence is released, collected and used for allele calling.

![Figure 6. Simplified drawing of the Taqman® genotyping assay. Modified from Life Technologies100.](image)

The Illumina Goldengate genotyping assay\textsuperscript{101} allows several loci to be examined in a single reaction through specific extension and amplification steps. It genotypes the targeted region directly on the genomic DNA and does not require prior PCR amplification of the genotyping target. The procedure of the assay works so that the DNA sample is activated for bind to paramagnetic particles. In the hybridization step the assay oligonucleotides, hybridization buffer, and paramagnetic particles are combined with the activated DNA. For each SNP, three oligonucleotides are designed; two are allele-specific oligos (ASOs) and the third hybridizes 1-20 bp downstream of the SNP site, and is the locus-specific oligo (LSO). All three oligos contain sequences that are recognized by a set of universal PCR primers. In addition, the LSO contain a unique sequence complementary to a particular bead type.
In this procedure, the assay oligonucleotides hybridize to the genomic DNA sample that is bound to the paramagnetic particles, upon which a polymerase extends the ASO. The polymerase fills the gap between the ASO and LSO and drops of when reaching the LSO. The nick between the extended sequence of the ASO and the LSO is sealed by a DNA ligase, forming PCR templates containing the genotype information present at the SNP site that will in the next step be amplified with universal PCR primers. The universal primers are labeled with different dyes. After thermal cycling the dye-labeled DNA products are hybridized to their complement bead type through their unique address sequence found in the LSO. At the end the Illumina BeadArray Reader is used to analyze the fluorescence signal from the Array Matrix or BeadChip. This technology is scalable from genotyping a single SNP to several thousands.

Sanger Sequencing
To determine the nucleotide order on a particular DNA fragment it is necessary to determine its sequence. A common technique that has been used for several decades is Dideoxy Sanger sequencing. The method is based on hybridization of a sequencing primer to a PCR product. In the next step deoxyribonucleotide triphosphates (dNTPs) and dideoxynucleotide triphosphates (ddNTPs) that lack a hydroxyl group at the 3’ end are incorporated. When the polymerase incorporates a ddNTPs the elongation of the synthesized strand is terminated, resulting in a range of fragments that differ in length by one base and are terminated at the point of incorporation of the base ddNTPs in question. Today, the ddNTPs are fluorescence labeled and the fragments are size separated using capillary electrophoresis (CE). The human genome was sequenced by fragmentation of DNA, which was cloned into bacterial and yeast vectors and sequenced using the Sanger technology.

Next Generation Sequencing
The last five years, the Sanger method is being replaced by Next generation sequencing (NGS) technologies, mainly for large-scale genome analysis.
NGS is a collective group of methods characterized by massively parallelization of the sequencing reactions, and therefore also called massively parallel sequencing (MPS)\textsuperscript{107}. There are different NGS platforms such as Illumina HiSeq/MiSeq, Life Technologies SOLiD and Ion Torrent/Ion Proton, and Pacific Biosciences RSII (Table 1). What separates NGS from Sanger sequencing is that NGS methods generate millions to billions of sequence in a single run. Most of the NGS technologies generate reads between 100-400. An exception is the Pacific Biosciences RSII with have a read length of 10-40 000 bp\textsuperscript{108}. Having an intention to lower the cost of DNA sequencing compared to the costs of standard dye-terminator methods\textsuperscript{109}. NGS platforms differs in the achievable read length, accuracy, reads per run, time per run, as well as the calculated cost per 1 million bases\textsuperscript{110,111}.

In this thesis two of the NGS sequencing platforms have been used, namely SOLiD sequencing and Pacific Biosciences RSII. The SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing platform was introduced in 2008\textsuperscript{112,113} and uses probes with dual-base encoding and ligation sequencing. The Pacific Biosciences RSII was introduced in 2009 and is a single-molecule real-time (SMRT) sequencing system\textsuperscript{114}.

In the first step DNA is randomly fragmented and universal nucleic acid adapters are ligated to the ends. This is generally denoted the sequencing library. The adapters are used to make the DNA fragments attach to a solid surface as well as to define the site of the sequencing start. These adapters look different depending on sequencing method. Linear adapters are used in SOLiD systems, while so-called bubble adapters are used in the Pacific Biosciences RSII system. In the second step emulsion-PCR is applied in the SOLiD and Ion systems, where a single enrichment bead and sequencing library fragment are combined inside an aqueous reaction bubble. A PCR is then performed inside the bubble resulting in clonal copies of the template on the surface of the bead. Beads with clonal copies of the DNA are attached to a glass slide. In the Pacific Biosciences RSII system, the biotinylated DNA polymerase binds to the bubble-adapted template and this complex is immobilized on the bottom of a zero mode wave-guide (ZMW) (small well). The sequencing is performed using DNA polymerase directed synthesis with fluorescent nucleotides. In the DNA synthesis each nucleotide is labeled by a fluorophore which gets cleaved of during the incorporation event. Pacific Biosciences RSII is based on observing the incorporation of nucleotides in real-time in parallel by recording the light pulses emitted during each incorporation event\textsuperscript{114}. In the SOLiD technology, sequencing is performed by ligation of fluorophore-labeled oligonucleotides that correspond to a specific base in the DNA molecule\textsuperscript{112,113}. 

28
<table>
<thead>
<tr>
<th>Platform</th>
<th>Company</th>
<th>Sequencing library</th>
<th>Support</th>
<th>Featured generation</th>
<th>Sequencing method</th>
<th>Detection method</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS FLX</td>
<td>454 Life Science, Roche</td>
<td>Linear adapter</td>
<td>Picotier plate</td>
<td>Emulsion PCR</td>
<td>Synthesis</td>
<td>Pyrosequencing</td>
</tr>
<tr>
<td>Genome</td>
<td>Solexa, Illumina</td>
<td>Linear adapters</td>
<td>Flow cell</td>
<td>Bridge PCR</td>
<td>Synthesis</td>
<td>Fluorophore labeled reversible terminator nucleotides</td>
</tr>
<tr>
<td>Analyzer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOLiD</td>
<td>Applied Biosystems</td>
<td>Linear adapters</td>
<td>Flow cell</td>
<td>Emulsion PCR</td>
<td>Ligation</td>
<td>Fluorophore labeled oligonucleotide probes</td>
</tr>
<tr>
<td>Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyzer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGA</td>
<td>Complete genomics</td>
<td>Circular adapters</td>
<td>DNA nanoball</td>
<td>Rolling circle</td>
<td>Ligation</td>
<td>Fluorophore labeled oligonucleotide probes</td>
</tr>
<tr>
<td>platform</td>
<td></td>
<td></td>
<td>arrays</td>
<td>amplification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PacBio RS</td>
<td>Pacific bioscience</td>
<td>Bubble adapters</td>
<td>Zero mode waveguide</td>
<td>Single molecule</td>
<td>Synthesis, Realtime</td>
<td>Phospholinked fluorophore labeled nucleotide</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Present investigation

Aim
The aim of this thesis is to study the association of genetic variation with development of cervical cancer, using both the candidate gene and GWAS approaches.

Paper I
_Evaluation of Fanconi Anemia genes FANCA, FANCC and FANCL in cervical cancer susceptibility_

Background
The FANC gene family consists of 13 FA complementation groups (FA-A, -B, -C, -D1, -D2, -E, -F, -G, -I, -J, -L, -M and –N). Disruption of the function in any of them causes Fanconi Anemia (FA)\textsuperscript{115}. FA is rare autosomal recessive DNA repair deficiency disorder, characterized by aplastic anemia and cellular hypersensitivity to DNA cross-linking agents\textsuperscript{116}. The FA patients are highly susceptible to different cancer types, including cervical cancer and other HPV associated tumors\textsuperscript{117}. The increased risk of cancer development has led to the suggestion that FA genes could play a role in recognition and repair of DNA damage. Genetic variation in DNA repair genes could affect identification and repair of DNA damage caused by the action of E6 and E7 HPV oncoproteins\textsuperscript{118}.

In paper I we are evaluating the association of 81 tagSNPs in three of the thirteen mentioned complementation groups of the FANC pathway: FANCA, FANCC and FANCL. The genes were selected based on earlier reports of their function or reported association with cervical cancer\textsuperscript{86}.

Results and discussions
TagSNPs were designed from the set of common SNPs genotyped in the Caucasian (CEU) population samples of the international HapMap project using the Tagger algorithm in the Haploview software\textsuperscript{119,120}. Genotypes were generated for 782 cases (CINIII and ICC) and 775 controls. A nominal asso-
ciation with cervical cancer was observed in FANCA and FANCC gene. However, none of these associations remained significant after Bonferroni correction for multiple testing, implying that the study gives no evidence for an association of the FA genes with cervical cancer in the Swedish population. In a previous study Wang et al. (2009) reported an association of variation at FANCA with cervical cancer in the Costa Rican population. The different results between our study and that by Wang could be due to haplotype differences between populations, resulting in different tagged variants. The Wang et al. results may also represent a false positive finding, since no replication of their findings has been published. TagSNPs are selected as to cover the genetic variation in the desired genes, but we cannot be certain that all SNPs are annotated in the HapMap project. Therefore the causal SNP could be missed and no association with cervical cancer will be found. This problem can be assessed with sequencing that would capture all genetic variation in the genes of interest.

In conclusion, the pathway still remains biologically relevant and we cannot exclude the possibility that more extensive genetic analyses of the FA genes would identify a contribution of the FA pathway to cervical cancer.

**Paper II**

*Contribution of TMC6 and 8 (EVER1 and EVER2) variants to cervical cancer susceptibility*

**Background**

The adjacent genes TMC6 and TMC8, also known as EVER1 and EVER2, have previously been associated with a rare autosomal recessive disease Epidermodysplasia Verruciformis (EV). The disease is characterized by a severe susceptibility to infections by cutaneous HPV of the beta genus and high risk of squamous cell carcinoma of the skin. The association between these two genes and the HPV sensitivity makes them interesting candidate genes in cervical cancer.

In paper II we are evaluating the association of 22 tagSNPs in the genomic region of TMC6 and TMC8 genes with susceptibility to cervical cancer. The genes were selected based on earlier reports of the functional consequence or reported association with persistent HPV infections and cervical cancer.

**Results and Discussion**

The 22 tagSNPs were selected using data from the HapMap CEU population in The International HapMap project and the Tagger algorithm in
Haploview\textsuperscript{119, 120}. The tagSNPs were genotyped in two Swedish cohorts, in total 2,989 cases and 2,281 controls\textsuperscript{117, 125}. A previous study by Wang et al. investigated the region of interest and reported an association of TMC6 and TMC8 genes with increased risk of cervical cancer in the Costa Rican population\textsuperscript{87}. In the Swedish population we identified two tagSNPs that were associated with cervical cancer, located in the intronic region of TNRC6C (SNP rs2290907), and an adjacent gene upstream of TMC6 and TMC8 gene (SNP rs16970849). We were not able to replicate the association of the two associated SNPs from the Costa Rican population in the Swedish cohort. These results may reflect the fact that we are comparing two ethnically distinct populations. In fact, only 13 tagSNP of the 22 studied here were common to both populations, and most of them vary in allele frequency between the studies.

Our observation confirms the involvement of TMC6 and TMC8 in the susceptibility of cervical cancer in the Swedish population as well. However, further studies of the region are needed to establish the full spectrum of genetic variation, as well as to identify functional variants.

**Paper III**

*Genome-wide association study of susceptibility loci for cervical cancer*

**Background**

GWAS have been successfully used to identify genetic associations across a wide range of traits and disease entities. In a GWAS the genome is searched for SNPs that occur more frequently in cases than controls.

In paper III we performed the first GWAS for cervical cancer, in order to further investigate the genetic basis of cervical cancer in the Swedish population.

**Results and Discussion**

After quality control the discovery data set included 632,668 SNPs that were genotyped in 1,034 cases and 3,948 controls\textsuperscript{117}. Testing each SNP individually using the log-additive model, we identified 55 SNPs to be associated with cervical cancer. All of them were found in the MHC region at 6p21.3, with the strongest association coming from SNP rs9272143 located between HLA-DRB1 and HLA-DQA1. This SNP is located within an intergenic region with no obvious functional elements. Conditioning upon SNP rs9272143 residual association was found at SNP rs2516448, adjacent to the MICA gene. This gene encodes a membrane-bound protein that acts as ligand for NKG2D activating immunoreceptor. Further statistical conditioning
lead to residual association found at an SNP (rs3117027) located in a pseudo gene of HLA-DPB2 that correlates with DPB1*0301, which has been found to be associated with increased risk of cervical cancer\textsuperscript{126}. However, further investigations are needed to determine if the association is driven by DPB1*0301. We could conclude that the three loci in the MHC region found to be associated with cervical cancer are independent of each other. One of the regions most strongly associated with cervical cancer is the MICA gene. The membrane-bound MICA protein is generally absent in most cells but is activated upon viral infections and reported to be expressed in epithelial tumors\textsuperscript{127}. The binding to MICA causes NKG2D to activate cytolytic response of gamma/delta T cells and natural killer cells against transfectants and tumor cells expressing MICA\textsuperscript{128}. Membrane-bound MICA therefore acts as a signal during the early immune response against infection or spontaneously arising tumors. The tumor cells on the other hand releases soluble MICA to damage the responsiveness of tumor antigen-specific effector T-cells, promoting tumor immune evasion as well as compromise the host resistance to infections\textsuperscript{127, 129}. The transmembrane domain (TMD) of the MICA gene contains a triplet repeat microsatellite polymorphism consisting of four, five, six, or nine repetitions of GCT (alleles designed A4, A5, A6, or A9, respectively) or five repetitions of GCT with one additional G nucleotide insertion (allele designed A5.1)\textsuperscript{130}. The insertion leads to a frameshift mutation that results in a premature stop codon (TAA), causing a truncation of 10 amino acids of the TMD as well as the hydrophobic 42-amino acid-long cytoplasmic tail. We found that the T (risk) allele of SNP rs2516448 is in perfect LD with the A5.1 mutation; as well as the A5.1 allele is associated with less membrane-bound MICA. The loss of 52 amino acids is believed to affect the membrane expression of the MICA protein. This may affect the ability of the cells to alert the immune system of HPV infection or neoplastic change, leading to impaired immune activation and increased risk of tumor development.

In an independent set of 1 140 case and 1 058 controls we replicated the association for two of the three SNPs. The effect for rs3117027 was just above the significance level. We also imputed the genotype alleles for the classic HLA alleles, and identified an association with cervical cancer with one HLA class I allele (B*0702) and five HLA class II alleles (DRB1*1301, DRB1*1501, DQA1*0103, DQB1*0603 and DQB1*0602), consistent with previously reported HLA association with cervical cancer in European populations\textsuperscript{72-75, 117, 131}. We could also identify DRB1*1301-DQA1*0103-DQB1*0603 haplotype to be associated with reduced risk of cervical cancer and B*0702 and DRB1*1501-DQB1*0602 haplotype with increased risk. Adjusting for the effect of the classic HLA alleles, the three novel loci remained associated with cervical cancer.

Furthermore, our GWAS associations have been followed up in several publications. It is known that multiple validations are of importance to con-
firm the GWAS findings and report the genotype-phenotype associations. Chen et al. performed an additional replication of the three novel loci identified in our first study as well as more extended genotyping of the MICA exon 5 microsatellite where the frameshift mutation (A5.1) is found. The C allele of SNP rs9272143 was found to have a protecting effect on cervical cancer and to be associated with higher expression of HLA-DRB\textsuperscript{132}. This SNP has been identified as a cis-expression quantitative trait locus (cis-eQTL)\textsuperscript{133}. That is loci regulating the levels of the mRNA expression\textsuperscript{134}. The T allele of SNP rs2516448 was associated with increased susceptibility to cervical cancer and the same association was shown with MICA-A5.1. The associations were consistent with our previous finding supporting the role of HLA-DRB1 and MICA in the pathogenesis of cervical cancer. Despite the findings in the GWAS no association was seen between the HLA-DPB2 variant SNP rs3117027 and cervical cancer\textsuperscript{132}. However, in a later study classical HLA-DPB1 alleles based on data from the MHC region were imputed. Allowing a systematic evaluation of the HLA-DP variants in the Swedish population. The strongest association found was for SNP rs3117027 that seems highly correlated with SNP rs3129294 that may have a putative regulatory function\textsuperscript{135}.

In conclusion, our GWAS study found risk of HPV-induced CIS to be associated with different polymorphic variations in several regions of the MHC. The results were also replicated in different cohorts, suggesting the MHC region as the main host genetic susceptibility factors to cervical cancer. Playing a role in the immune recognition of HPV-infected or –transformed cervical cells causing increased risk of tumor development. Still, further studies using for instance genome-wide sequence-based association analysis could be of importance, to fully define the inherited basis of cervical cancer.


d Paper IV

Characterization of the TMC6 and TMC8 genes associated with cervical cancer.

Background

Previous findings have shown that polymorphisms in TMC6 (EVER1) and TMC8 (EVER2) are associated with increased risk of cervical cancer, or susceptibility to HPV or other genital infections\textsuperscript{87, 136}. As previously mentioned these two genes were first been associated with Epidermodysplasia Verruciformis (EV)\textsuperscript{121-123}. In paper IV we used a variety of different genetic technologies to characterize the genomic region harboring the TMC6 and
TMC8 genes, in order to study their association with cervical cancer in more details.

Results and Discussion

Using long-range PCR in combination with SOLiD sequencing, the distribution of SNPs across the TMC6 and TMC8 region were first identified in pools of cases and controls\textsuperscript{137}. Large allele frequency differences were seen between pools for a number of SNPs, and the 10 SNPs with the largest difference were selected for replication. Replication of the top ten SNPs using TaqMan genotyping assay resulted in that only one SNP (rs2926778) showed a significant allele frequency difference between cases and controls. This SNP is located between TMC6 and the upstream gene, TNRC6C, and has not been included in the previous tagSNPs analysis by Castro \textit{et al} (2013)\textsuperscript{136}. However the polymorphism is located in the intronic region and the putative functional consequence is unknown. The SNP rs66468151 instead of three different clusters in the Taqman genotyping (homozygote for A allele, heterozygote for AB alleles and homozygote for B allele) showed 5 different genotype clusters. Using predesigned Taqman CNV assays we then identified that 3% of the cases have a copy number variation (CNV) in the examined region. This corresponds to the region with the SNP with the highest allele frequency difference in the pooled sequencing. It is therefore likely that the failure to validate the top ten SNPs is due to the complexity of the DNA sequence, such as repeats or other structural variants in this region.

To increase our understanding of the region, we attempted to sequence the region using Sanger sequencing\textsuperscript{103}. This resulted in complete sequences for some individuals, but a gap in individuals with the previously noted CNV. We therefore used a second sequencing approach, based on hybridization capture of 2-3 kb genomic fragments, and the long-read single-molecule DNA sequencing of Pacific Biosciences RS II. A \textit{de novo} assembly of the sequencing reads was conducted resulting in a complete contig for the region, providing a more detailed understanding of the nature of the variation. Together with the new hg19 we improved the assembly of TMC6 and TMC8 region and were able to determine the nature of the variation. In the region where we previously have indications of a CNV, we found a 21 bp sequence that was be repeated 15 times in the reference sequence and 18 times in the \textit{de novo} assembly. Knowing the exact region of the CNV we could identify 39 different alleles in our cervical cancer cases and controls. The alleles range in size from 0 to 42 repeated units of the 21bp sequence. No significant (p>0.05) allele frequency difference was seen between cases and controls when testing all alleles. However, when comparing the group allele denoted “0” versus the rest resulted in significantly lower frequency of this allele in the cases as compared to controls. (p=0.05) This 21 bp sequence is not found elsewhere in the human genome and does not contain any identi-
fied binding motifs, indicative of its function. At present, we are not able to suggest a functional role of this repeat. Also, there might exist SNPs and other types of variations in the region flanking the repeat, within the “0” allele group that could be of functional importance. Further sequencing of additional individuals using the capture and Pacific Biosciences method is needed to address these questions.

In conclusion, our study have resolved the TMC6-TMC8 genomic region previously associated with cervical cancer, and identified an association for the SNP rs2926778 located between TMC6 and TNRC6C gene with cervical cancer. We have also identified structural differences in the DNA sequence of the TMC6 gene. Lack of repeat units was found to be associated with increased risk of cervical cancer. However, further studies are needed to understand the nature of these associations.

Figure 8. Schematic drawing of the study methods used.
Conclusion and future perspectives

Cervical cancer is the fourth most commonly diagnosed cancer, and the fourth leading cause of cancer death, among women worldwide. In 2012, 528 000 new cancer cases and 266 000 deaths due to cervical cancer was registered. Around 85% of new cancer cases occurred in the less developed countries, where also almost nine out of ten (87%) cervical cancer deaths occurred. This thesis is focusing on evaluating the genetic variation in the human genome using the GWAS as well as different candidate genes, which have previously been associated with cervical cancer or other disorders that are of importance in infection by different HPV’s as well as the progression to cervical cancer.

The first, second and the fourth paper are all based on previous associations found with cervical cancer. The first study could not find any indication of association in the investigated Fanconi anemia (FA) genes with cervical cancer in the Swedish population. The second paper presents association of variation in the region of two genes; TMC6 and TMC8, earlier associated with the HPV-associated disease Epidermodysplasia Verruciformis (EV) as well as increased risk of HPV associated cervical cancer. In the fourth paper a new polymorphism was associated with cervical cancer in the same region as studied in paper three. Upon sequencing of the complex genomic region of TMC6-TMC8, a CNV was identified, and the lack of this repeat was found to be associated with increased risk of cervical cancer. In the third paper three novel loci in the major histocompatibility complex (MHC) region were found to be associated with the disease, highlighting the importance of variation in this regions for the risk of developing cervical cancer. In conclusion, the results presented point to the involvement of several genetic components in the development of cervical cancer. The genetic risk factors that are identified to date have inadequate effect size, and only explain a portion of the genetic susceptibility, which implies that further investigations are needed to fully understand the contribution of genetic factors to this disease.

Previous studies have emphasized the importance of genetic predisposition in the disease. Magnusson et al. demonstrated a family clustering of cervical cancer. He showed that the relative risk in developing cervical cancer was higher among biological mothers compared to the adoptive mothers to probands. The same was noted for the biological sisters versus adoptive sisters\(^69\). In 2000 Lichtenstein et al. demonstrated that the risk of developing
cervical cancer was 0% explained by genetic factors, 20% was explained by shared environment and 80% was explained by non-shared environmental factors. In 2006 a Danish group reported family clustering of cervical cancer in both mono- and dizygotic twins. They suggested that family factors such as genes or and shared environment are important in the development of cervical cancer\textsuperscript{138}. Taking the information together we are getting different suggestions of the genetic importance in the development of cervical cancer. A couple of years ago it was believed that GWAS would help us uncover and explain some of the genetically diseases. Many of the GWAS of complex disease have indeed pointed out regions with associated genes in breast\textsuperscript{139-141}, colon\textsuperscript{142, 143} and prostate cancer\textsuperscript{144, 145} so there is no doubt that GWAS have resulted in interesting candidate regions for some of the examined phenotypes. Still, most of the variation that has been identified so far have small effect and only explain a small part of the heritability. Are we missing something in our way of studying the human genome, are there still variants that we have not discovered?

In 2008 the involvement of structural variation, such as copy number variation (CNV), was demonstrated in schizophrenia\textsuperscript{146}. These findings have lead to an interest in evaluating the potential importance of structural variation in several common human complex diseases. Next generation sequencing (NGS) technologies have given us an opportunity to study the human genome in even more detail. Perhaps in the future these studies will help us answer the question of genetic importance in the cervical cancer.

Could it be so that cervical cancer is an immunodeficiency disease? Several studies\textsuperscript{72, 73} including our GWAS\textsuperscript{147} have emphasized the involvement of the immune system genes in the disease. In particular the MHC region, that plays an important role in presenting foreign antigens to the immune system. Failure in activating and developing an effective cell-mediated immune response to take care of the HPV infection could result in a persistent infection as well as increased risk of cervical cancer development\textsuperscript{148}.

It is of importance to remember that cervical cancer is a mixture of an infectious disease and a cancer. It differs from the majority of cancer since viral infection of the HPV is an essential cause. Knowing this it is of high importance to prevent the infection. Today two vaccines have been available for almost a decade, which have proven so far to have a long lasting effect. However, the issue still remains that the long-term protection is unknown; therefore long-term follow-up studies are necessary. Another thing is that the screening should never be allowed to get in the background due to the available vaccination. It is of importance to continue monitoring the prevalence of different HPV types in infected women, due to that the vaccines only protect against HPV16 and 18 and we do know that there are several other oncogenic HPV types that could affect the cancer development.
Acknowledgements

Tack till min handledare Ulf Gyllensten för att jag fått doktorera hos dig. Tack för all stöd, alla diskussioner och all vägledning. Du är otroligt duktig, kunnig och påläst, men samtidigt en jordnära och fin människa.


Tack till Sara, Dijana & Mi för det vi en gång var, för alla fester, diskussioner & vår resa till Milano. Ett extra tack till Mi för alla uppmuntrande ord under mitt skrivande. Till min Ghazal, min iranska själsfrände. Jag är glad att jag har mött dig under min doktorandtid för jag vet att jag har hittat en vän för livet.

Tack till mina gamla som nya vänner, för all stöd under dessa år & alla peppande ord de sista månaderna.

Tack till mina svärföräldrar Vesna & Miro, min mans syster Ivana med respektive för alla stöttande ord.

“Vi ne birate svoju obitelj. Ona je Božji dar vama, kao što ste i vi njima.”
Desmond Tutu (A vi ste moj najljepši dar od Boga)

References

31. Sustained immunogenicity and efficacy of the HPV-16/18 AS04-adjuvanted vaccine: up to 8.4 years of follow-up. in *Hum Vaccin Immunother*, Vol. 8 390-7 (United States, 2012).


63. Cancer, I.C.o.E.S.o.C. Cervical carcinoma and reproductive factors: collaborative reanalysis of individual data on 16,563 women with


79. Gostout, B.S. et al. TAP1, TAP2, and HLA-DR2 alleles are predictors of cervical cancer risk. in Gynecol Oncol, Vol. 88 326-32 (United States, 2003).


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1100

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-248484