UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Technology

Ksenia Chloe Bartlett

Genetic Susceptibility Factors of High-Risk Human Papillomavirus (HPV)

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:

Associate Professor, Triin Laisk (PhD)

Co-supervisor:

Specialist, Mariann Koel (MSc)

Tartu 2023

Genetic Susceptibility Factors of High-Risk Human Papillomavirus (HPV)

Abstract:

Cervical cancer is a significant health concern linked to high-risk human papillomavirus (HPV) infection. This study aimed to explore the genetic susceptibility factors associated with high-risk HPV infection using a genome-wide association study (GWAS) approach. The analysis included 3445 cases and 10467 controls, utilizing women with positive high-risk or potentially high-risk HPV test results as cases and those without as controls.

The GWAS analysis identified a locus for high-risk HPV infection within the MHC region on chromosome 6p21.3. To further elucidate the specific human leukocyte antigen (HLA) alleles associated with high-risk HPV infection, the signal was mapped, revealing *HLA-DQB10603, HLA-DRB1*1301, HLA-DRB1*13, HLA-DQA1*0103, HLA-DQB1*0602,* and *HLA-DRB1*13 as the most significantly associated alleles.

In addition, the study identified 15 diagnosis codes with significantly different prevalences in cases and controls, indicating associations between high-risk HPV infection and viral agents, abnormal findings in female genital specimens, malignant neoplasm of the cervix uteri, and other sexually transmitted diseases.

Several limitations were encountered, including the absence of a specific code for high-risk HPV infection, which affected sample identification and size. The lack of a replication cohort necessitates further validation in independent datasets. Furthermore, the transient nature of most HPV infections posed challenges in distinguishing controls with no high-risk HPV infection history.

In conclusion, this research highlights the need for larger GWAS studies to investigate the genetics of high-risk HPV infection comprehensively. Discovering additional genetic susceptibility factors can enhance prevention strategies and contribute to advancements in personalized medicine.

Keywords: cervical cancer, high-risk human papillomavirus infection, genetic susceptibility factors, genome-wide association study, human leukocyte antigen (HLA), MHC region.

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics; B570 Obstetrics, gynecology, andrology, reproduction, sexuality

Kõrge riskiga inimese papilloomiviiruse (HPV) geneetilised riskitegurid

Lühikokkuvõte:

Emakakaelavähk on pahaloomuline kasvaja, mida põhjustab nakkus inimese papilloomiviirusega (HPV). Käesoleva töö eesmärgiks oli ülegenoomse seoses uuringuga (GWAS) leida nn kõrge riskiga HPV tüvede geneetilisi riskifaktoreid. Analüüsi kaasati 3445 juhtu ning 10467 kontrolli, kusjuures juhtudeks olid naised, kellel on HPV testimisel leitud nn kõrge või võimaliku kõrge riskiga HPV tüvesid, samas kui kontrollideks olid naised, kelle HPV testi tulemused olid negatiivsed.

GWAS uurings leiti, et kõrge riskiga HPV infektsiooniga on seotud lookus MHC piirkonnas 6. kromosoomil (6p21.3). Et täpsemalt kaardistada, millised inimese leukotsüüdi antigeeni (HLA) alleelid on seotud kõrge riskiga HPV infektsiooniga, teostati seoseanalüüs HLA alleelidega, ning leiti, et *HLA-DQB10603*, *HLA-DRB11301*, *HLA-DQA10103*, *HLA-DQB10602*, and *HLA-DRB11*3 alleelid on statistiliselt oluliselt seotud.

Lisaks leiti uuringu käigus 15 haiguskoodi, mille esinemissagedus juhtude ja kontrollide seas oli oluliselt erinev, sh diagnoosikoodid, mis on seotud viirusinfektsioonidega, naissuguelunditest pärit materjali leidude hälvetega, emakakaelavähiga, ning sugulisel teel levivate infektsioonidega.

Uuringu puuduseks võib pidada seda, et kõrge riskiga HPV infektsioonil puudub eraldi diagnoosikood, mis raskendab juhtude ja kontrollide identifitseerimist. Lisaks oleks kirjeldatud leida vaja valideerida edasistes uuringutes. Meeles peab pidama ka seda, et enamasti HPV infektsioon laheneb iseenesest, mis raskendab tõeliselt infektsioonivabade kontrollisikute identifitseerimist.

Kokkuvõttes näitab meie uuring, et HPV infektsiooni geneetiliste riskifaktorite leidmiseks on vaja suuremahulisi ülegenoomseid uuringuid ning uute geneetiliste riskifaktorite leidmine võib aidata praeguseid ennetusmeetmeid tõhusamaks muuta ning personaalmeditsiini edendada.

Võtmesõnad: emakakaelavähk, kõrge riskiga inimese papilloomiviiruse infektsioon, geneetilised riskifaktorid, ülegenoomne seoseuuring, inimese koesobivusantigeen (HLA), peamise koesobivuskompleksi (MHC) regioon

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetria; B570 Sünnitusabi, günekoloogia, androloogia, reproduktsioon, seksuaalsus

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TERMS, ABBREVIATIONS, AND NOTATIONS

CC	Cervical cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
EstBB	Estonian Biobank
GWAS	Genome-wide Association study
HLA	Human Leukocyte Antigen
HPV	Human papillomavirus
LOINC	Logical Observation Identifiers Names and Codes
MAF	Minor allele frequency
MHC	Major histocompatibility complex
PC	Principal components
pheWAS	Phenome-wide association study
SNP	Single Nucleotide Polymorphism
STIs	Sexually transmitted infections

INTRODUCTION

Cervical cancer is one of the most prevalent cancers among women, accounting for many deaths worldwide. Infection by high-risk human papillomavirus (HPV) has been established to be one of the primary causal agents of cervical cancer. However, not all individuals that have been subject to high-risk HPV infections develop malignancy, suggesting that genetic-susceptibility factors play a role in the pathogenesis of cervical cancer. Cervical cancer and HPV infection are both complex traits, which means their risk and progression are impacted by genetic, environmental, and lifestyle factors.

Nowadays, genome-wide association studies (GWAS) are commonly used to determine the genetic determinants of complex traits and diseases. While extensive research has been done on the genetics of cervical cancer, studies concerning the genetic susceptibility factors of high-risk HPV infection have mostly been candidate-gene studies. Also, these studies had small sample sizes leading to a decreased statistical power.

In this study, we will do extensive research on the genetic susceptibility of high-risk HPV infection by annotating the results from a genome-wide association study of high-risk HPV status, characterizing the associations in the HLA region in more detail by visualizing and interpreting results from an HLA-allele association study and determining the phenotypes associated with high-risk HPV infection status in a phenome-wide association study. Hopefully, our research will contribute to the existing database of high-risk HPV infection genetics.

1 LITERATURE REVIEW

The human papillomavirus (HPV), which is considered to be one of the causes of infection and malignancy in the uterine cervix, is a highly varied virus having more than 200 genotypes with different virulence [14]. HPV is pronounced to be one of the leading causes of uterine cervix cancer, with an estimated 99.7% of cervical cancer cases being caused by persistent genital high-risk HPV infection [48]. Cervical cancer is the fourth most prevalent cancer in women globally and the third most common type among women between the ages of 15 and 44 in Europe, with an estimated 58,169 new instances of cervical cancer in 2020 [2]. Infection with recognized high-risk human papillomavirus subtypes (HPV 16, 18, 31, 33, 34, 35, 39, 45, 51, 52, 56, 58, 66, 68, and 70) as well as environmental or lifestyle factors have been shown to contribute to the development of CC [17]. Despite the central role of the HPV in the development of cervical cancer, host genetic factors also play a vital role in determining the susceptibility to and outcome of the infection [1], and recently, several genes associated with susceptibility to cervical cancer have been identified in genome-wide association studies (GWAS). At the same time, studies exploring the genetic determinants of HPV infection susceptibility and outcome remain sparse and underpowered, with most of them being candidate gene association studies. Thus, we lack knowledge of the genetic susceptibility factors of high-risk HPV infection.

1.1 Uterine Cervix and HPV Infection

The uterine cervix is a part of the uterus located at its lower end; anatomically, it comprises an internal and external layer of tissues called the endocervix and ectocervix. The uterine cervix is cylindrical in appearance, and its length is 4 cm, while its width is 3 cm; this part of the uterus connects the uterine cavity to the vagina [19]. The two layers meet in a so-called transformation zone as the ectocervix's outer lining of squamous cells combines with the endocervix's columnar glandular cells [19] (Figure 1.1).

The main biological role of the uterine cervix is to control the timing of sperm entry and also to form a barrier for microorganisms. The uterine cervix separates the lower and the upper parts of the genital tract, thus forming a barrier that prohibits the infection of the upper genital tract by pathogens and plays an essential role in female reproductive health [1]. Namely, the cervix is the main infection site for several sexually transmitted infections (STIs), such as chlamydia, gonorrhea or *Mycoplasma genitalium*, and HPV.

The uterine cervix (Figure 1.1) is susceptible to HPV infection that may cause pre-cancerous lesions in this region [17]. The vaginal part of the uterine cervix is covered with stratified squamous epithelium, while its cervical canal part comprises long cells with non-granular cytoplasm [13]. However, mainly during HPV infection, the outer columnar epithelium that lines the endocervical canal of the ducts and the glands is transformed into the squamous epithelium.

Before cancer cells begin their formation in cervical tissues, the cells undergo abnormal alterations known as dysplasia (Figure 1.2) [38]. There are several forms of dysplasia [38]. One of those forms is mild dysplasia, often known as a low-grade intraepithelial lesion (LSIL) [38]. Another kind of dysplasia is moderate or severe, often known as a high-grade intraepithelial lesion (HSIL) [38]. LSIL and HSIL may or may not develop into malignancy [38].

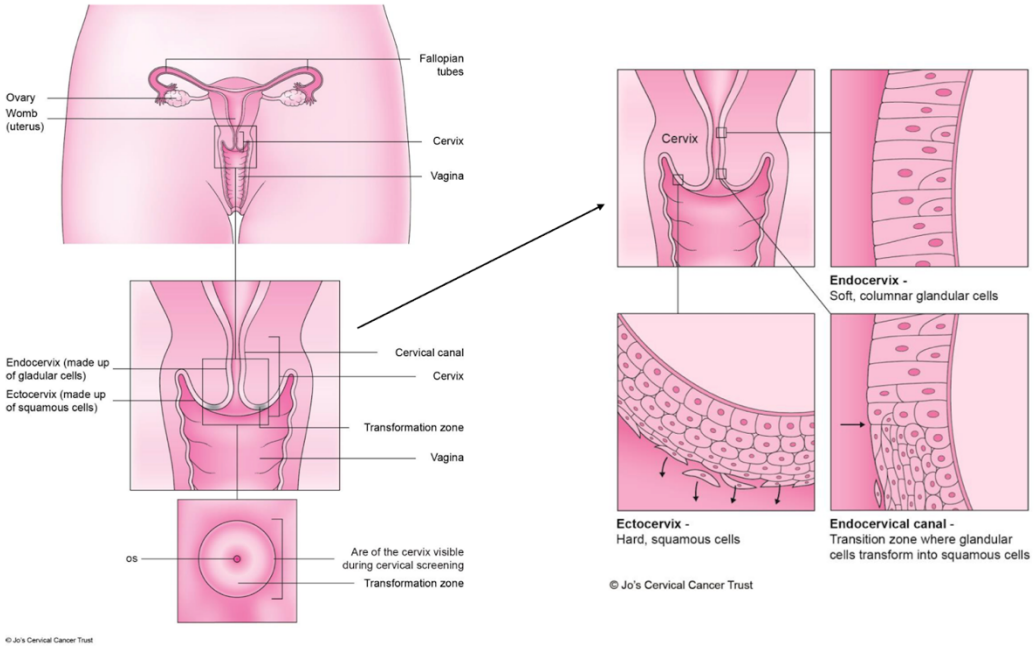

Figure 1.1: The location and anatomy of the uterine cervix by Jo's Cervical Cancer Trust (2020) [4]. The following diagram shows the location of the uterine cervix and the internal and external layers it is composed of (endocervix and ectocervix).

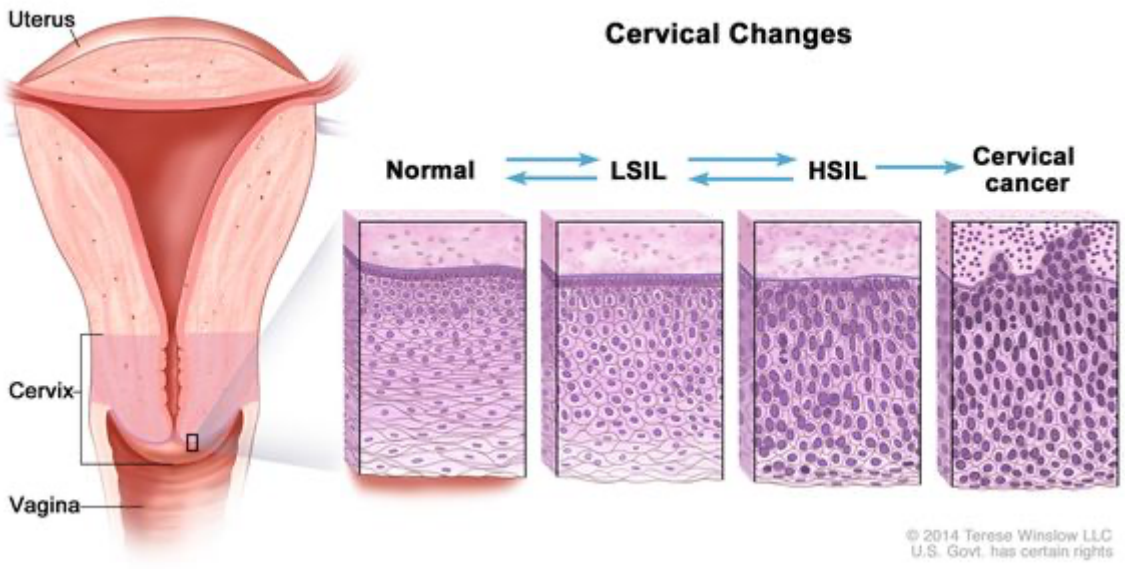

Figure 1.2: Cervical changes before the formation of cancer cells [38]

In addition to the cervix, different HPV strains can also infect other epithelia and skin, causing different problems, such as viral warts, genital warts, and other types of genital and oral cancer. HPV infections are typically spread by direct skin or mucous membrane contact with an infected lesion [25]. In most cases, genital HPV infection is obtained through sexual contact; however, non-penetrative genital contact, oral-genital contact, and manual-genital touch are other potential modes of transmission [25]. Furthermore, genital HPV infection can be transferred perinatally from infected moms to neonates via the mouth and upper respiratory tract [25]. Personal skin-to-skin contact is also significant for nongenital HPV infection, while fomite transmission from damp surfaces is likely a major source of infection for plantar warts [25]. Autoinoculation can spread both genital and nongenital infections to new areas [25].

HPV infection can occur in histologically normal tissue and is only detectable through molecular approaches identifying viral DNA [32]. Most infections are asymptomatic and resolve independently; just 10% are predicted to last more than two years [37]. Only HPV infections that are permanently detectable are linked to the development of high-grade squamous intraepithelial lesions and cancer [37].

HPV is among the most prevalent STIs, yet it is not always associated with sexual behavior or promiscuity [33]. The precise frequency of HPV infection is unknown because HPV-related infections are not reportable STIs, and most cases are asymptomatic or subclinical [33]. Most HPV infections (90%) resolve naturally [33]. HPV 16 is the most common among

high-risk HPV strains; after HPV 16, the second most carcinogenic strain of HPV is HPV 18, as it is responsible for 12% of squamous cell carcinoma (SCC) as well as 37% of adenocarcinoma (ADC) [12]. HPV16 and -18 are responsible for over 80% of cervical malignancies [33]. These high-risk subtypes of HPV strains are one of the most crucial causes of cancer, especially in women; according to one estimate, various strains of this virus cause CC, which leads to the death of 236,000 women annually. Interestingly, human squamous epithelial cells can be immortalized in vitro by high-risk HPV strains 16/18 [33].

1.1.1 Cervical Dysplasia and Cancer

HPV infection is one of the prominent causes of uterine cervix cancer [18], with 99.7% of cervical cancer cases being caused by persistent genital high-risk human papillomavirus (HPV) infection [48]. Cervical cancer is the 4th most common cancer among women worldwide, with an estimated 604,127 new cases and 341,831 deaths in 2020, making it one of the leading causes of cancer-related mortality among women worldwide [2]. In Europe, there are roughly 58,169 new instances of cervical cancer detected each year (estimates for 2020), making it the ninth most common cause of cancer among women [2]. Cervical cancer is the third most prevalent type of female cancer in Europe among women between the ages of 15 and 44 (Figure 1.3) [2].

The infection by known high-risk sub-types of human papillomavirus (HPV 16, 18, 31, 33, 34, 35, 39, 45, 51, 52, 56, 58, 66, 68, and 70) (Table 1.1) along with environmental or lifestyle factors such as smoking, multiple sexual partners, high parity, obesity, early menopause, co-infection with *Chlamydia trachomatis*, herpes simplex virus type-2 (HSV2), or human immunodeficiency virus (HIV) are known to lead to the initiation of cervical cancer development [17]. The host's genetic makeup also plays a role in cervical cancer development by influencing whether the infection is susceptible to treatment or if it persists, ultimately leading to the formation of cervical cancer: The rate at which the tumor develops can also be influenced by host genetics [1]. Heritability estimates for cervical cancer from previous family-based studies ranged from 13 to 64%, and recent extensive genome-wide association studies (GWAS) have reported several genetic susceptibility factors for cervical cancer [1]. These will be discussed in more detail in Chapter 1.2.2.

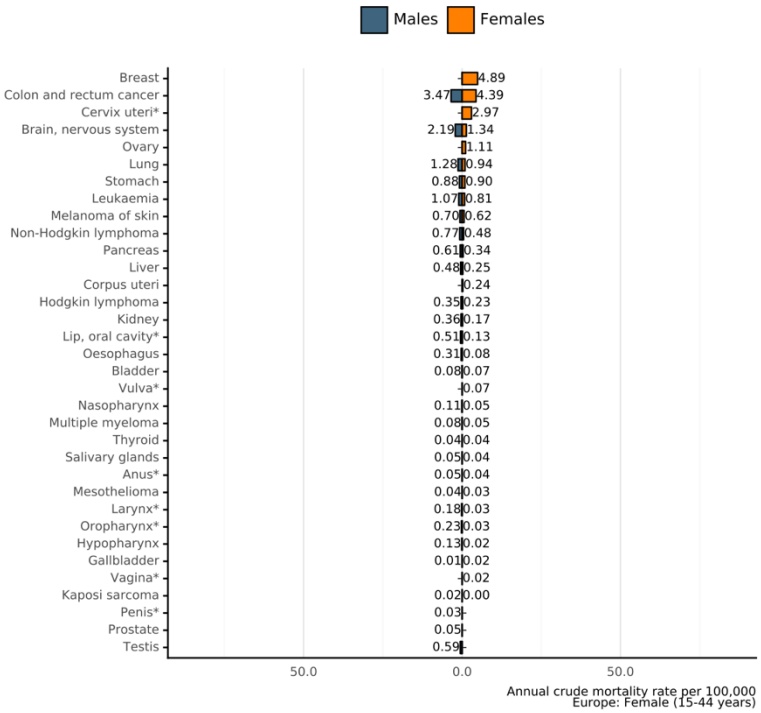

Figure 1.3: Comparison of HPV related cancers incidence to other cancers among men and women 15-44 years of age in the World (estimates for 2020) [2]

Table 1.1: HPV types and their classifications

Classification	HPV types
Carcinogenic (high-risk)	16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, 82
Possibly carcinogenic (intermediate risk)	26, 53, 66, 67, 70
Unknown risk	30, 34, 69, 85, 86, 97
Low-risk	6, 11, 39, 42, 43, 44, 54, 61, 70, 72, 81, CP6108

In most cases, HPV infects the basal epithelial layer. The vast majority of these infections are temporary and are eliminated by the immune system within a few years [39]. However, as the red arrows show (Figure 1.4), 10-20% of infections survive latently, leading to disease development [39]. The resulting lesion is termed central intraepithelial neoplasia (CIN) and is graded according to severity [39]. Low-grade squamous intraepithelial lesions (LSIL) eventually progress to high-grade squamous intraepithelial lesions (HSIL), which can lead to invasive cancer [39]. Despite tumor reduction in response to the first therapy, as shown by the green arrows (Figure 1.4), most cases with latent infection impede full clearance of the viral infection, resulting in lesion recurrence [39].

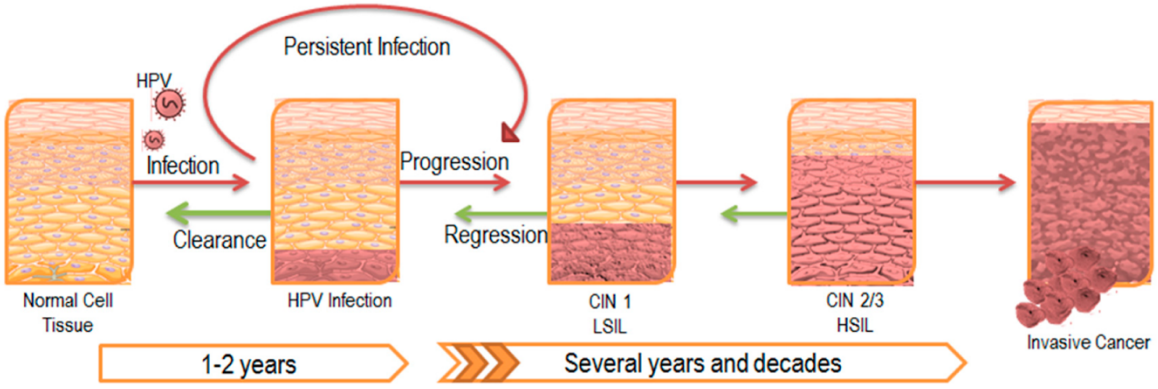

Figure 1.4: HPV Infection and Associated Disease Progression [39]

1.1.2 Prevention of Cervical Cancer

Several preventive strategies have been established to reduce cervical cancer incidence and mortality: well-organized cervical screening programs, widespread good quality cytology, and self-testing at home, followed by Papanicolaou "pap smear testing" and PCR-based tests for HPV typing. In the ensuing decades, the advent of HPV vaccination may also significantly lower the incidence of cervical cancer [22]. In many developed countries, vaccinations against the most common high-risk HPV types (Cervarix, Gardasil, or Gardasil 9) are offered [3]. However, because the vaccine was first launched in the 2000s, it was not feasible to prove definitively that it decreases cervical cancer incidences - the ultimate objective of the immunization program - until recently [51]. One study from Sweden discovered that HPV vaccination was responsible for a 63% reduction in cervical cancer incidence [51].

In Estonia, until 2021, women aged 30-55 with health insurance (roughly 90%) were asked to participate in an organized CC screening every five years, with the option of receiving a

free Pap smear regardless of screening history [23]. Since 2021, the target population for CC screening has been expanded to include women over 65 and women without health insurance and the HPV test has been introduced as the primary test. In the case of a positive test result, the same specimen will be subjected to a liquid-based cytology test (LBC-based Pap test) [23]. In Estonia, the HPV vaccine Gardasil 9 is used for vaccination, which protects against nine forms of HPV (6, 11, 16, 18, 31, 33, 45, 52, and 58) [40].

Despite the screening and vaccination strategies, in Estonia, according to data from the last five years, cervical cancer is the second most prevalent gynecological malignancy, and it is estimated that 160 women are diagnosed with cervical cancer per year [23]. Thus we need better tools to motivate participation in screening programs or increase vaccine uptake. Here, personalized genetic risk prediction might become useful. Still, before we can implement that, we need a better understanding of the host genetic risk factors of both HPV infection and cervical cancer development.

1.2 Studying the Genetics of complex traits and Diseases

As outlined above, cervical cancer and HPV infection are complex traits by nature, meaning their risk and development are influenced by genetic, environmental, and lifestyle factors. Nowadays, GWAS are commonly used to determine the genetic determinants of complex traits and diseases.

1.2.1 GWAS overview

GWAS is an approach that involves the study and profiling of millions of genetic variations and their alleles to identify and locate genetic variants more than likely to be associated with a disease or trait. In the previous decade, development and advancement in human genome research have made GWAS one of the most reliable and inclusive techniques for understanding the genetic architecture of human diseases (Figure 1.5) [11].

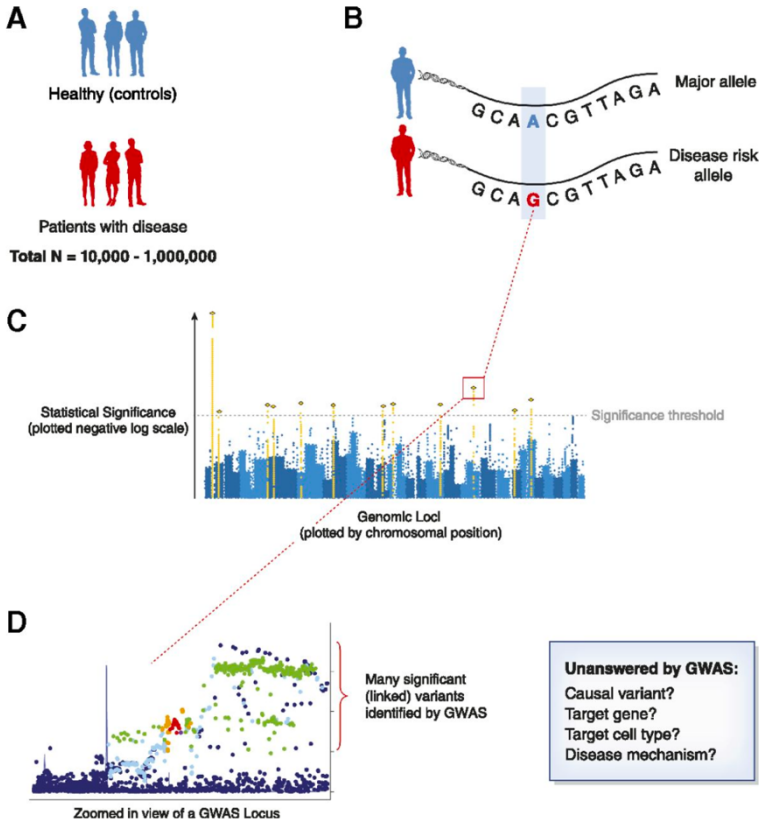

Figure 1.5: The Principle of a Genome-wide Association Study (GWAS) by Lin & Susztak (2020) [1]**.** A case-control design is commonly used in binary trait GWAS: disease cases (A) and healthy controls (B). Following that, the frequency of single nucleotide polymorphisms (SNPs) in these groups is compared [43] (B). The most statistically significant SNPs can be displayed in the form of a "Manhattan plot" [43] (C). A locus zoom graphic (D) shows how many genetic variants within one locus may imply a link to disease development [43]. It is critical to highlight that the GWAS does not reveal the causative variant, the target gene, the target cell type, or the disease mechanism [43].

In principle, GWAS is an instrument that expands researchers' ability to identify and explain genetic risk factors for several widespread diseases. As far as the uses and application of

GWAS are concerned, this method provides information about the genetic basis of the development of specific phenotypes. Overall, in combination with multiple follow-up methods, GWAS is a good tool for studying the basic biology of a trait or a disease.

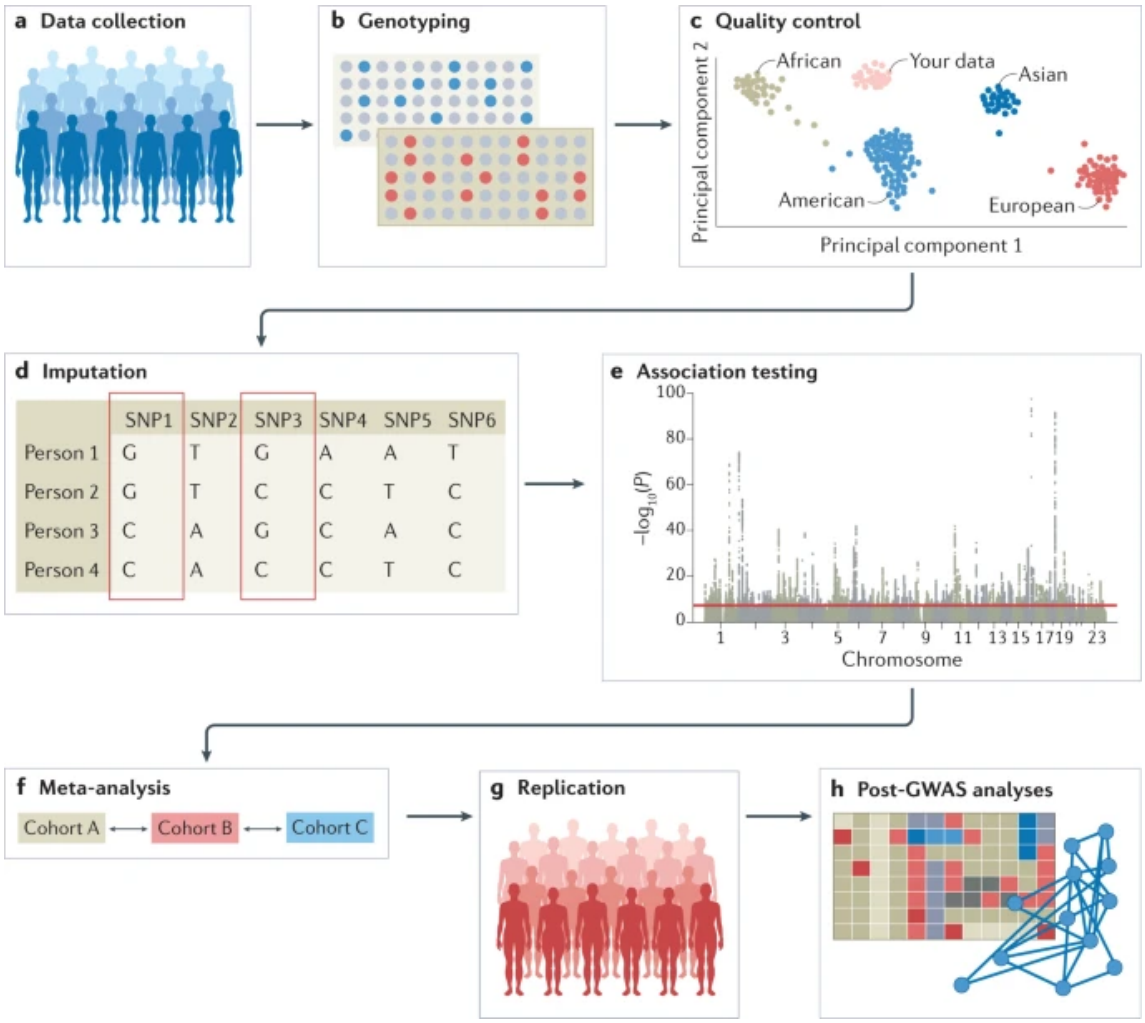

Figure 1.6: Overview of the steps for conducting GWAS from data collection to potential post-GWAS analyses [19]

To conduct a GWAS study, data can be acquired from study cohorts or from biobanks or repositories that have available genetic and phenotypic information (Figure 1.6) [19]. To obtain genotype information, microarrays can be used to capture common variants, or next-generation sequencing methods can be put into use for whole-genome sequencing (WGS) or whole-exome sequencing (WES) [19]. Due to cost, array genotyping is usually preferred.

Quality control in Genome-Wide Association Studies (GWAS) encompasses wet and dry-laboratory steps. Wet-laboratory steps involve processes such as genotype calling and DNA

switches, while dry-laboratory steps involve actions taken on called genotypes, including the removal of poor-quality single nucleotide polymorphisms (SNPs) and individuals, detection of population strata within the sample, and calculation of principal components [19]. The grouping of individuals based on genetic principal components can be visualized in Figure 1.6 [19].

To increase the number of evaluated variants and enhance cost-effectiveness, untyped genotypes can be imputed using information from matched reference populations, such as those available through the 1000 Genomes Project or TopMed [19]. In this case, genotypes of SNP1 and SNP3 are imputed based on the genotypes of other directly assayed SNPs, thereby expanding the scope of variants that can be assessed [19].

Once the genotypes are obtained, appropriate genetic association tests are performed for each variant using an applicable model, depending on the phenotype type and research question. Examples of such models include additive, non-additive, linear, or logistic regression [19]. Confounding factors, including population strata, are accounted for, and multiple testing is controlled [19]. The output is thoroughly examined for unusual patterns, and summary statistics are generated [19].

Standardized statistical pipelines are employed to aggregate results from numerous smaller cohorts, ensuring consistent analysis and interpretation [19]. Replication of findings, whether internally within the study or externally in an independent cohort, is employed to validate the results [19]. For external replication, the independent cohort must have a similar ancestral background and should not include any individuals or family members from the original discovery cohort [19].

The results obtained from GWAS serve as valuable input for subsequent analyses aimed at understanding disease biology better. In silico analysis of GWAS summary statistics often utilizes data from external sources [19]. This includes tasks such as in silico fine-mapping, SNP-to-gene mapping, gene-to-function mapping, pathway analysis, genetic association analysis, Mendelian randomization, and polygenic risk prediction [19]. Furthermore, functional hypotheses generated from GWAS can be further evaluated through experiments such as CRISPR or massively parallel reporter assays or validated in human trait/disease models [19].

By following these steps and conducting additional analyses, GWAS contributes to our understanding of the genetic basis of traits and diseases, facilitating advancements in disease research and personalized medicine.

1.2.2 Genetics of cervical cancer

In the case of the development of cervical cancer in women, various research studies have used the GWAS approach to identify susceptibility genes. Based on the data obtained from the Swedish cancer registry, disease clustering in families is prominent in the case of cervical cancer in women, revealing a high probability of cervical cancer in female siblings [9]. Similarly, some other genomic studies on the development of cervical cancer in women have found that this disease can develop even without HPV infection. In this case, the tumors have specific molecular pathology [20]

In this regard, one study has estimated that the proportion of genetic factors involving susceptible genes for the development of cervical cancer ranges between 27 to 36% (Magnusson et al., 2000). In line with this, GWAS has identified specific loci and potential candidate genes responsible for the development of cervical cancer. For instance, the human leukocyte antigen (HLA) locus in the chromosome 6p21.3 region [17], as well as non-HLA signals on chromosomes 2q13 (*PAX8*), 5p15.33 (*TERT-CLPTM1L*), and 17q12 (*GSDMB*) ([17], [10])..
Similarly, other studies have identified novel risk loci associated with cervical cancer or dysplasia. The GWAS conducted on a sample of Chinese people identified two risk loci: 4q12 (rs13117307, *EXOC1*) and 17q12 (rs8067378, *GSDMB*). GWAS using data obtained from the Japanese Biobank, also identified two unique gene variants responsible for CC located at *INS-IGF2* (rs150806792) and *SOX9* (rs139991990) [17].

The association in the HLA region has also been mapped to specific alleles. *HLA-DRB1**1301 and *DQB1**0603 alleles were found to be associated with decreased risk of cervical cancer ([27], [29], [30]), and more broadly, the *HLA-DRB1**1301–*HLA-DQA1**0103–*HLA-DQB1**0603 haplotype conferred the strongest protection against cervical cancer [26].

The continuous discovery of significant regions on the genome, especially in the HLA region, signifies the decisive role of genetic factors in the development of CC; however, less is known about the genetic risk factors of high-risk HPV infection.

1.2.3 Genetics of HPV Infection

Genetic association studies attempt to identify a link between the presence of a disease or condition and genetic variation to discover potential genes or genomic regions that impact how susceptible an individual is to that disease [24]. Since a large number of HPV-infected individuals never develop symptoms, it is difficult to determine whether a given SNP allele or genotype increases the risk of an HPV infection in a fraction of cases (HPV-infected subjects) and controls (individuals without the infection) [24]. Unfortunately, most gene association studies to identify genetic variants and susceptibility to HPV infections have been carried out utilizing the candidate gene technique, which confines the study to one or a few genes and is based on a previously established hypothesis regarding the significance of a selected gene [24]. These studies are often designed as case-control studies, in which cases and controls are identified first, then the genetic differences between the two groups are discovered later [24]. SNPs in genes linked with immune response, apoptosis, DNA repair, and within the HLA region have been evaluated and proposed as predictors of susceptibility to persistent HPV infection [24]. The most frequently documented genetic risk factor is allelic variation within the HLA region [24].

HLA is a group of highly polymorphic genes on human chromosome 6 that encode cell-surface proteins vital in immune regulation [24]. By delivering HPV-derived peptides to T-cells, HLA class I and II cell surface molecules regulate the host's immune system [28]. T-cell activation may differ based on the HLA allele polymorphism [28].

One study evaluated the association between three single nucleotide polymorphisms in the *IL10* promoter and clearance of low- or high-risk HPV infection in a cohort of 226 largely HIV-1–infected adolescent females [24]. It was found that among immunocompromised people (HIV-1 seropositive and CD4+ 500), the GCC haplotype in the *IL10* promoter was associated with a decreased clearance of high-risk HPV16-like, HPV18-like, and any high-risk type, but not with low-risk HPV type [31]. These findings suggested that IL10 variants influence the clearance of infection with high-risk HPV types because higher levels of *IL10* may impair the production of inflammatory cytokines such as *IL-2, TNF-*, *IL-4*, *IL-6*, and *IL-12*, which are involved in *TH1-TH2* immunoregulation and immunity against HPVs [24].

A case-control study of 161 cases and 257 controls of Brazilian women found that women with no HPV-related cancer who carried the alleles *DRB1*1503*, *DRB1*0395*, and *DQB1*0602* were more likely to test positive for HPV [32].

Another study conducted among 172 Mexican women observed that the *HLA-DRB1*07* allele had been associated with viral clearance [24]. Conversely, the variant *HLA-DQB1*0501* was associated with increased susceptibility to HPV reinfection, and the allele *HLA-DRB1*14* was associated with a putative protective factor for the development of cervical cancer [34].

In one study conducted among 541 Canadian women, the allele *HLA-G*01:01:01* was associated with an increased risk of alpha groups 1 and 3 infections (alpha group 1 LR-HPV cervical species; group 2 HR-HPV cervical species; and group 3 LR-HPV vaginal species), while no allele or genotype was associated with HPV persistence [35]. The effect of HLA-E and HLA-G polymorphisms on HPV infection susceptibility and persistence was investigated in 636 female university students in Montreal [36]. Persistent HPV-16 and HPV types from species 2, 3, 4, and 15 infections were more prevalent in women with *HLA-G*01:01:02* [36]. HLA-E variants, on the other hand, were not linked to the probability of HPV infection acquisition or persistence [36].

A recent study looked at the influence of variation on *HLA-DRB1* and *DQB1* HLA- alleles connected to the clearance of six HPV strains (HR-HPV-16, -18, -31, -33, -45, and -58) in 276 Colombian women [46]. It was found that the HLA allele/haplotype relationship with HPV infection clearance associated with the infecting HPV type, in accordance with the specific viral epitopes displayed, and thus, while *DRB1*12:01:01G* favored HPV-16 and HPV-45 clearance, it restricted HPV-18, HPV-31, and HPV-58 elimination.

In another study among 517 Nigerian women, *DQA1*01:02* and *DQA1*02:01* were found to be strongly associated with prevalent but not persistent HR-HPV infections, and four haplotypes (*A*30:01-DQA1*05:01*, *B*07:02-C*07:02*, *B*07:02-DQA1*05:01*, and C*07:02-DQA1*05:01) were found to be significantly associated with persistent cervical HR-HPV infections [47].

At the same time, all of the studies have been relatively small in size, with sample sizes ranging from 172 to 636, meaning that, most likely, we still lack robust evidence to link HPV infection risk and outcome to specific genetic susceptibility factors.

2 THE AIMS OF THIS THESIS

This thesis aims to characterize the genetic susceptibility factors of high-risk HPV infection status in the Estonian Biobank. To achieve this, the following sub-aims were set:

1. To annotate the results from a genome-wide association study of high-risk HPV status using the FUMA platform

2. To characterize the associations in the HLA region in more detail by visualizing and interpreting results from an HLA-allele association study

3. To determine the phenotypes associated with high-risk HPV infection status in a phenome-wide association study setting

3 EXPERIMENTAL PART

3.1 Materials and Methods

All the analyses described here were conducted using the Estonian Biobank data under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs) and data release N05. Analyses involving individual-level data were carried out by researchers with the necessary approval (Mariann Koel and Triin Laisk). My role was to annotate the results from a genome-wide association study of high-risk HPV status using the FUMA platform, characterize the associations in the HLA region in more detail by visualizing and interpreting results from an HLA-allele association study, and determine the phenotypes associated with high-risk HPV infection status in a phenome-wide association study setting.

3.1.1 Estonian Biobank (Study Design and Participants)

The cohort of the Estonian Biobank is an adult resident community sample drawn from volunteers aged 18 or older [5]. The current participant count, which is over 200,000, reflects a sizable portion, or 20%, of the entire adult population of Estonia, subsequently making it the perfect choice for population-based studies [5].

Information on high-risk HPV infection was extracted from electronic health records from the Estonian Biobank by the STACC, which is the leading machine learning and data science company in Estonia, using Logical Observation Identifiers Names and Codes (LOINC) codes for HPV testing. The following LOINC codes were used for extracting individuals who had been tested for a high-risk HPV type: L-2712, 61372-9, 61373-7, 71431-1, and 49896-4.

Data was further organized and cleaned by Mariann Koel. As a result, we identified:

- 10586 women who were tested but did not have any positive test results
- 985 women had a positive test, but the HPV subtype was unknown
- 3091 women who had a positive test for a specific HPV subtype
    - 209 of them had one or several positive results for LOW-risk subtypes- 53, 82, 42, 44, 54, 43, 61, 39

- 2882 of them had one or several positive results for high-risk HPV subtypes

For the following analyses, all women, who had at least one positive high-risk or potentially high-risk test result, were used as cases. Women, who had been tested, but received a negative result, were used as controls.

3.1.2 Quality Control and association testing

All EstBB participants have been genotyped at the Genotyping Core Lab of the Institute of Genomics, University of Tartu, using the Illumina Global Screening Array v1.0 and v2.0 [1]. Illumina GenomeStudio v2.0.4 was used to genotype the samples and generate PLINK format files. If a participant's call rate was < 95% or if the sex defined based on heterozygosity of the X chromosome did not match the sex in phenotype data, they were not included in the analysis [1]. Before imputation, variants were subject to filtering by the following criteria: call rate < 95%, HWE p-value < 1e-4 (autosomal variants only), and minor allele frequency (MAF) < 1% [1]. Using GSAMD-24v1-0_20011747_A1-b37.strand, all variants in the human genome b37 were altered to be from the TOP strand.RefAlt.zip files from the website https://www.well.ox.ac.uk/wrayner/strand/ [1]. Eagle v2.3 software [6] was used for pre-phasing (the number of conditioning haplotypes Eagle2 employs when phasing each sample was set to: --Kpbwt=20000), and Beagle v.28Sep18.79339 was used for imputation (effective population size ne=20,000) [1]. Population-specific imputation reference of 2297 WGS samples was used [7]. Genotyping, quality control, and imputation were carried out centrally by the Genotyping Core Lab and Bioinformatics Core Lab of the Institute of Genomics, University of Tartu [1].

Association analysis was carried out using REGENIE software implementing a mixed logistic regression model [8], using the year of birth and ten genetic principal components (PCs) as covariates in step I [1]. The final analysis included 3445 cases and 10467 controls with genotype data available. Triin Laisk did the association testing, and the resulting summary statistics were used in follow-up analyses.

3.1.3 Annotation of high-risk HPV infection status GWAS using the FUMA platform

FUMA (v1.4.0) [22] was used to annotate the GWAS signals using the SNP2GENE function.

FUMA is a web-based digital platform that uses information related to the molecular biology of genes from several resources and makes functional annotation of GWAS results easy and

smooth. FUMA is an important innovation as it resolves the main issue of pinpointing possible causal variants in GWAS since the hits are mostly concentrated in non-coding or intergenic regions [21]. For the same reason, post-GWAS annotation is recommended as it helps pick up the most probable causal variants and genes. One option is using multiple resources for the post-GWAS annotation; nonetheless, using these resources can prove to be time-consuming; additionally, this approach does not have the integrated visual aid necessary for the effective interpretation of data [21].

Using FUMA resolves all these issues as it uses digital means to assist in the functional annotation of GWAS results and gene prioritization and provides interactive visualization. Some of the most effective aspects of FUMA are that it accommodates positional mapping of candidate genes in the locus, as well as mapping based on quantitative expression trait loci (eQTL) or chromatin interactions. Additionally, the platform enables researchers to perform gene-based analyses, as well as pathway and tissue enrichment analyses [21]. The information obtained from FUMA is beneficial in developing hypotheses, which are verifiable using functional experimentation conducted to verify biological mechanisms. FUMA's most significant working principle is that it uses 18 biological databases, for example, Ensembl, GTEx, and GWAS catalog, and uses tools to process and interpret the summary of GWAS results [21].

The main function of FUMA is called SNP2GENE; during this process, FUMA performs annotation of SNPs using their biological functionality; afterward, SNPs are mapped to genes in accordance with following the following type of SNPS information: positional, eQTL, and chromatin interaction (Figure 3.1) [21]. This process results in identifying prioritized genes using the following mapping techniques: positional, eQTL, and chromatin interaction mapping.

After completing the analyses, the results are projected in graphical forms; for this purpose, FUMA uses either interactive tables or plots. This tool also allows downloading the tabulated information in plain text files. For the graphs, this tool provides the option of downloading them in the form of HD images using different formats [21]. FUMA is a comprehensive bioinformatics tool that incorporates most of the features of other bioinformatics tools. For instance, it incorporates LD calculation, variant annotation, gene-based test/gene-set analysis, and visualization [21]. Based on this description of FUMA, one can firmly assert that it is one of the most effective and user-friendly bioinformatics tools available for annotating GWAS results and therefore was also used in this thesis.

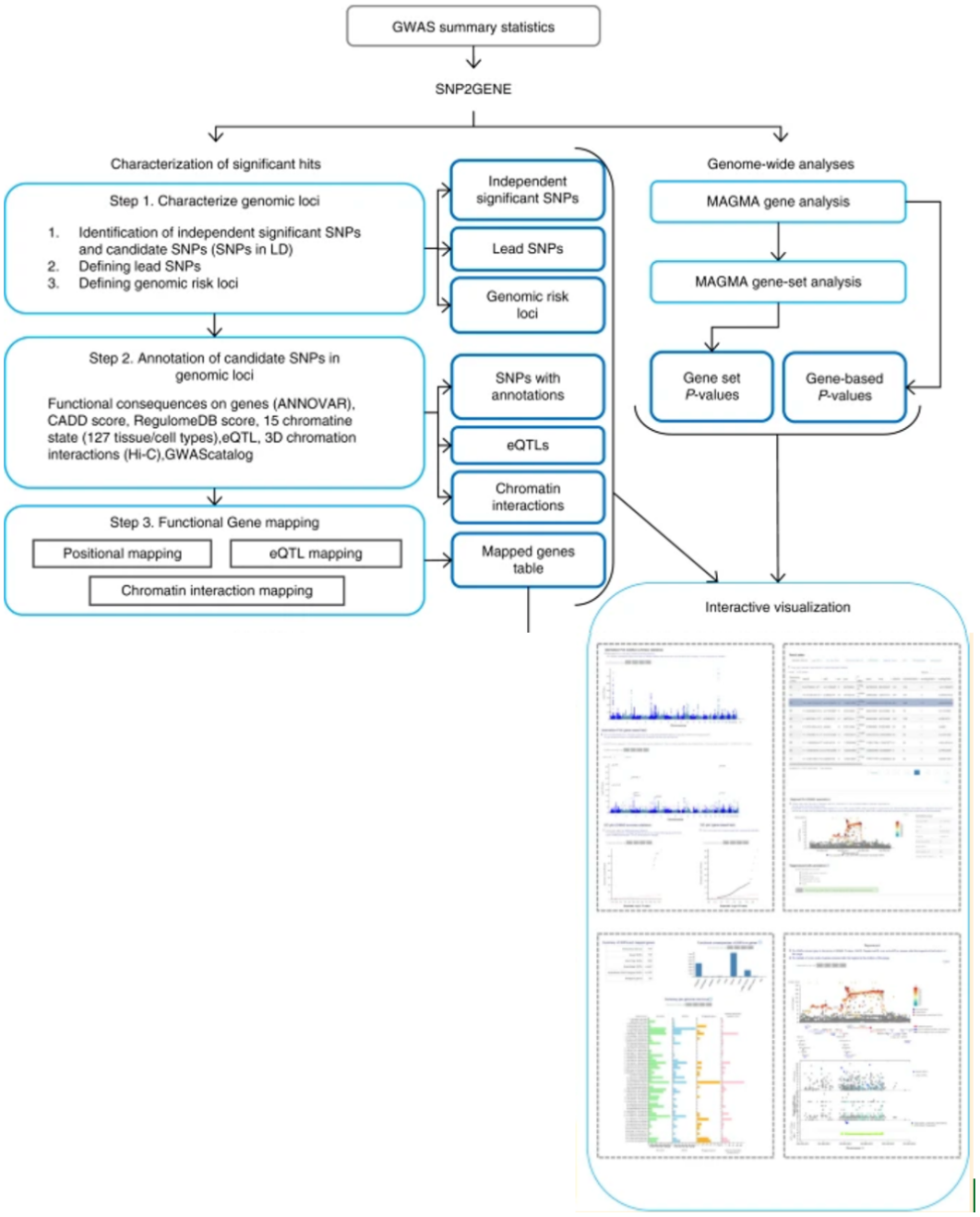

Figure 3.1: Overview of the steps and function (SNP2GENE) of FUMA for the annotation of GWAS signals [44]

3.1.4 Association testing of high-risk HPV infection and HLA alleles (HLA analysis)

The signal that was detected by annotating the GWAS results using FUMA alone does not tell us much about the specific HLA alleles that are associated with our trait of interest (high-risk HPV infection). To overcome this, a statistical method has been developed to facilitate the fine mapping in the HLA region and "pick apart" the GWAS signal we see in our original analysis. There are several software platforms to do this, but at the Estonian Biobank the SNP2HLA v1.0.3 tool [20] has been used to impute classical alleles at class I (*HLA-A*, *-B*, *-C*) and class II (*-DPA1*, *-DPB1*, *-DQA1*, *-DQB1*, and *-DRB1*) loci. The SNP2HLA technique utilizes long-range linkage disequilibrium between HLA loci and SNP markers to impute amino acid polymorphisms and alleles, enabling the determination of HLA variance in available GWAS data (Figure 3.2) [16]. As an imputation reference, a merged reference of EstBB WGS [41] and Type 1 Diabetes Genetics Consortium samples [42] was used [1]. We tested for association between the alleles and high-risk HPV infection in the EstBB using the SAIGE software with the LOCO option. Imputed data on alleles was used (two- and four-digit) in the MHC class I genes (*HLA-A*, *HLA-B*, *HLA-C*) and classical MHC class II genes (*HLA- DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, *HLA-DPB1*) for 3325 cases and 10153 controls in the EstBB, who had the corresponding data available.

The results were sorted by p-value to see which HLA alleles are most significantly associated with our trait (high-risk HPV infection). For this, we used the R-studio [45] to sort by p-value using the p-value column.

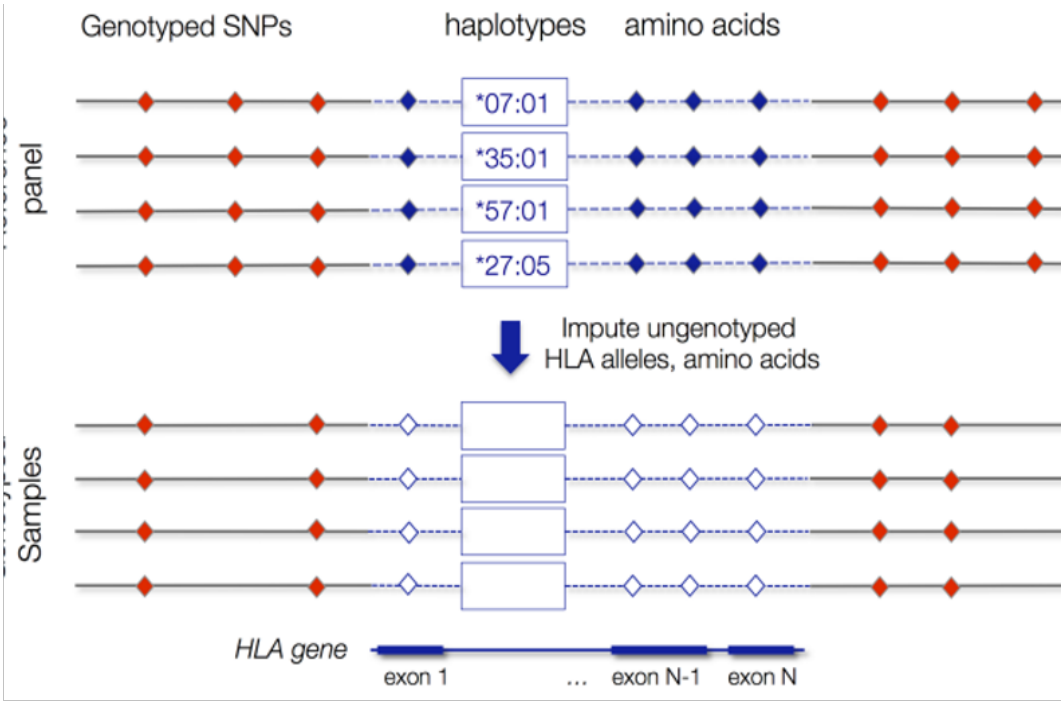

Figure 3.2: SNP2HLA imputation procedure [16]. SNPs in the MHC, traditional HLA alleles at class I and class II loci, and amino acid sequences matching the 4-digit HLA types at each locus are all represented in the reference panel (top). The reference panel imputed classical alleles and their corresponding amino acid polymorphisms for a data set comprising genotyped SNPs throughout the MHC (bottom) [16].

3.1.5 Analysis of phenotypes associated with high-risk HPV status in EstBB

The individual-level data in the EstBB was used to proceed with an analysis to uncover international classification of diseases 10th revision (ICD10) diagnosis codes associated with the high-risk HPV infection diagnosis [3]. This could provide information on the potential health outcomes of high-risk HPV infection, as well as highlight diagnoses that most commonly occur together with a high-risk HPV infection, providing input for further analyses. A logistic regression framework assessed the association between high-risk HPV infection and ICD10 codes, adjusting for age and ten genetic PCs [3]. Due to the EstBB including a high number of relatives and because the inclusion of relatives may exaggerate the association statistics, all first- and second-degree relatives (genetic relatedness cut-off value 0.2) in pairwise comparisons were removed, maintaining cases, if possible, to avoid the loss of power [3]. This yielded 3375 cases and 6032 controls for the analysis [3]. To find statistically significant relationships, the Bonferroni correction was used (number of tested ICD major codes - 2001, adjusted p-value threshold – 2.5×10^{-5})) [3]. The PheWas library 0.99.5-4 was used to visualize the results. R 3.6.3 [45] was used for all analyses [3].

3.2 Results

3.2.1 Genome-wide association study analysis

The analysis identified one locus for high-risk HPV infection, with one locus significantly associated with high-risk HPV infection ($P < 5 \times 10^{-8}$).

According to FUMA, there is only one signal, and it is a common variant rs3892710 (l p=2.17x10^{-9}, OR=1.27 (1.19-1.35) in the human major histocompatibility complex (MHC) located on the short arm of chromosome 6 (6p21. 3; Figure 3.3). Effect allele frequency in cases (0.86) is bigger than in controls (0.83). Alleles *HLA-DQB1* and *HLA-DQA2* are near the signal (Figure 3.4).

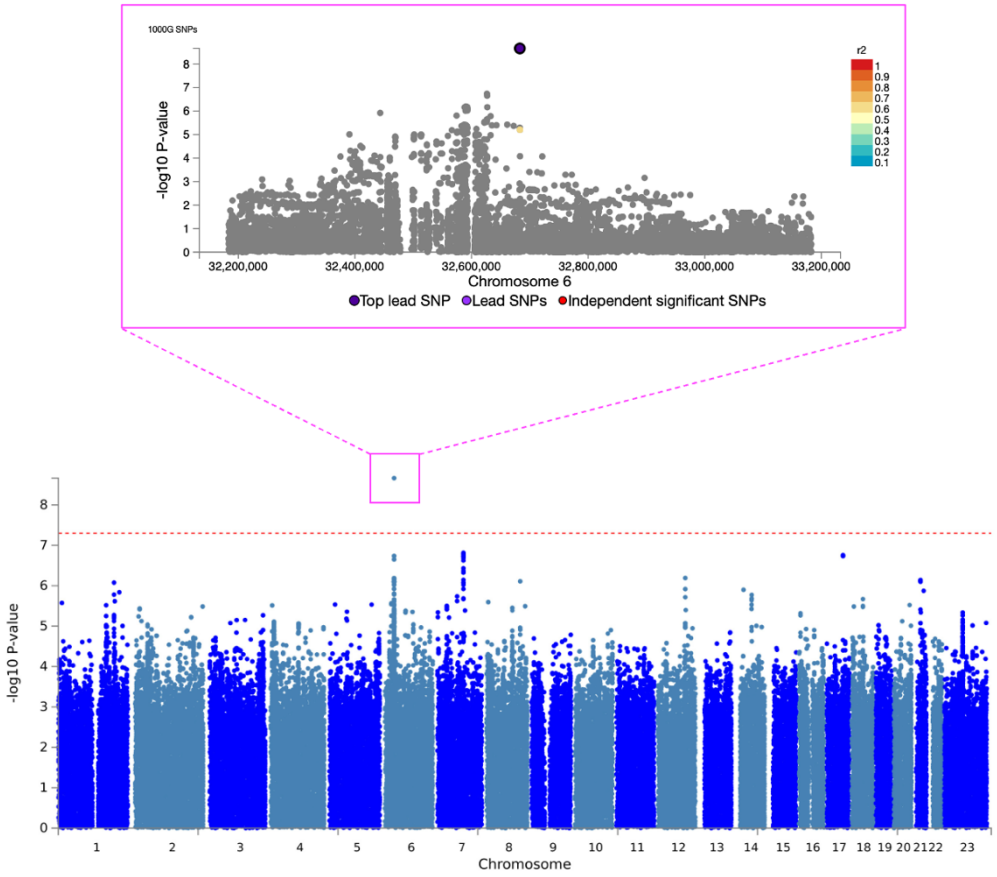

Figure 3.3: Manhattan Plot (GWAS summary statistics) and regional plot for high-risk HPV infection GWAS analysis signal on chr6 [44]. The y-axis of the Manhattan plot reflects -log10(P-values) for the association of variants with high-risk HPV infection. The horizontal dashed line indicates the threshold ($P < 5 \times 10^{-8}$) for genome-wide significance. The lead variant of a genome-wide significant locus on chromosome 6 is depicted in a regional plot.

3.2.2 Association testing of high-risk HPV infection and HLA alleles

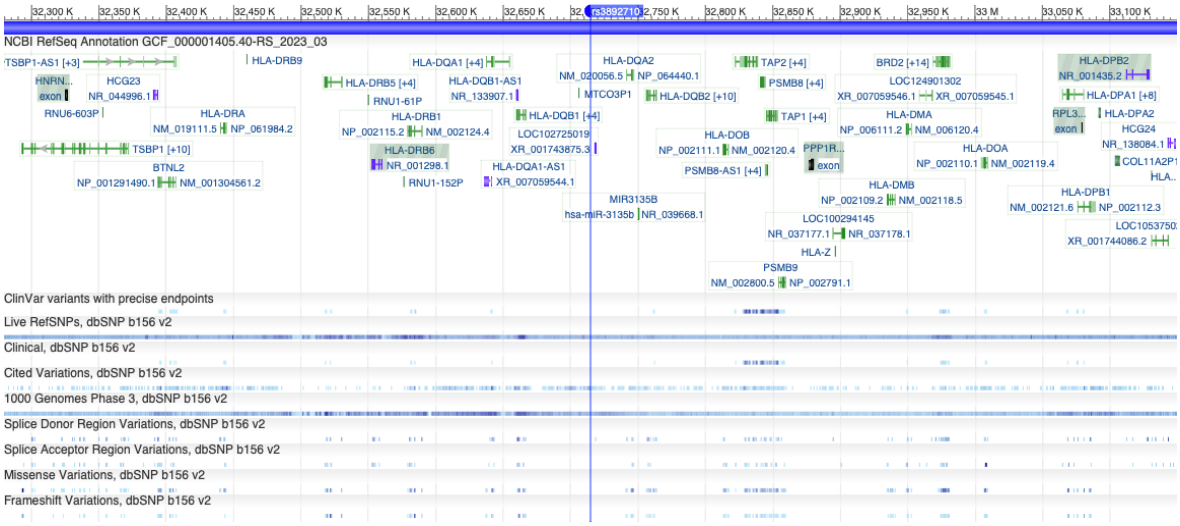

Figure 3.4: The location of the lead signal rs3892710 and the genes in the locus visualized in the dbSNP database browser [50].

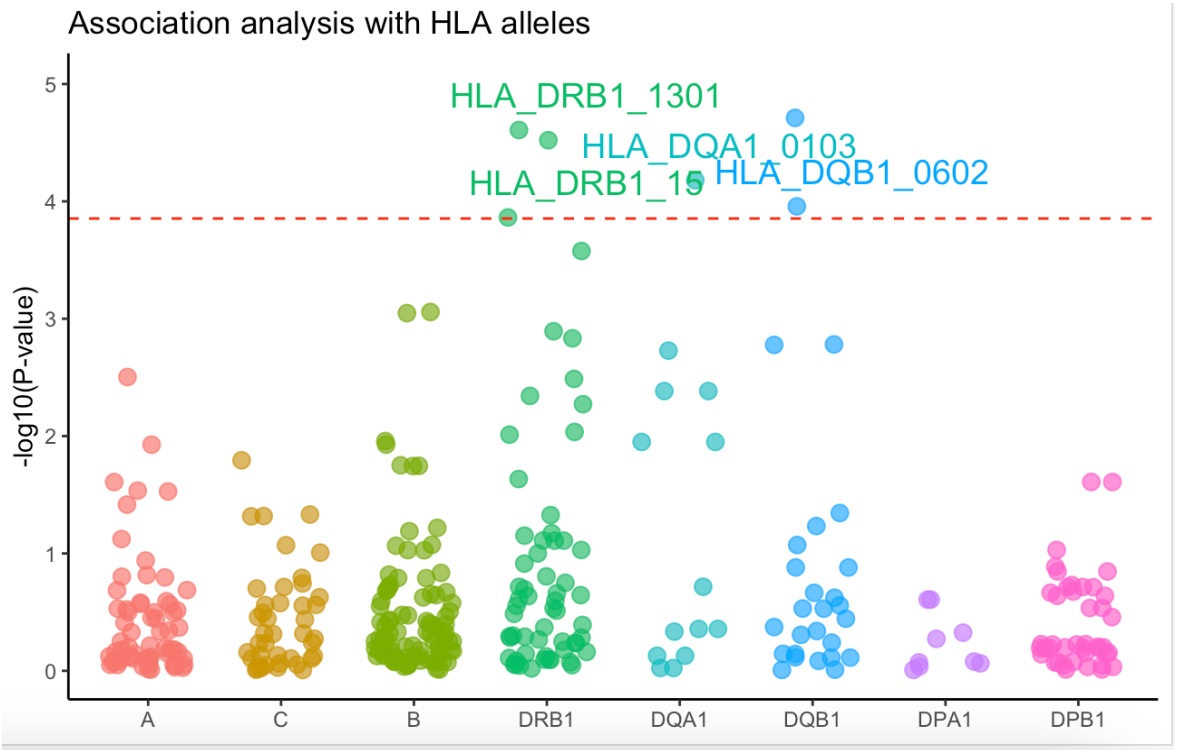

Figure 3.5: HLA alleles associated with high-risk HPV infection [44]. The y-axis shows the -log10 P-values from the analysis of 3325 cases and 10153 controls in the EstBB using SAIGE [1]. The red dashed line represents the p-value threshold adjusted for the number of tested alleles ($p<1.4 \times 10^{-4}$).

Since high-risk HPV infection shows an association signal in the HLA region, the imputed HLA allele dataset in EstBB was used for further mapping the high-risk HPV infection association signal in the HLA region (Figure 3.6). ***HLA-DQB1*0603*** ($p=1.9 \times 10^{-5}$, OR=0.79 (0.71-0.88)), ***HLA-DRB1*1301*** ($p=2.5 \times 10^{-5}$, OR=0.80 (0.72-0.89)), ***HLA-DRB1*13*** ($p=3.0 \times 10^{-5}$, OR=0.83 (0.76-0.91)), ***HLA-DQA1*0103*** ($p=6.6 \times 10^{-5}$, OR=0.81 (0.73-0.90-)), ***HLA-DQB1*0602*** ($p=1.2 \times 10^{-4}$, OR=1.18 (1.08-1.28),) and ***HLA-DRB1*15*** ($p=1.4 \times 10^{-4}$, OR=1.18 (1.08-1.28) alleles passed the significance threshold after correcting for multiple testing.

These results are in line with previous studies in cervical cancer - *HLA-DRB1*1301* and *DQB1*0603* alleles are associated with decreased risk of cervical cancer ([27], [29], [30]) and more broadly, the *HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603* haplotype conferred the strongest protection against cervical cancer [26].

3.2.3 Associated phenotypes

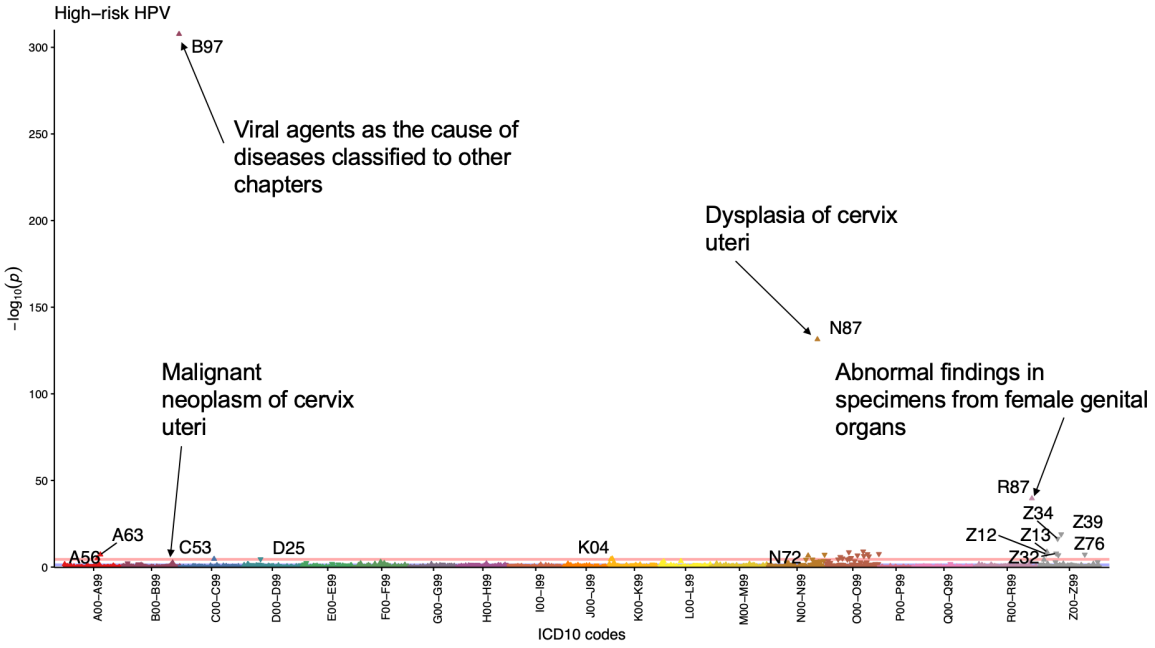

Figure 3.6: Disease codes associated with a diagnosis of high-risk HPV infection in the Estonian Biobank [45]. Each triangle represents one ICD-10 main code, while different colors correspond to different chapters. The direction of the triangle is an illustration of effect direction- upward pointing triangles show increased prevalence of the diagnoses codes in high-risk HPV infection cases. The Bonferroni corrected threshold for statistical significance is represented by a pink line.

The significantly associated disease codes in the pheWAS analysis are consistent with what is known about the etiopathogenesis of the condition, validating the used phenotype definition. In our analysis, women diagnosed with high-risk HPV infection have significantly more diagnoses of viral agents as the cause of diseases classified to other chapters (B97, OR=12.4 (11.1-13.8). The association we observe is most likely driven by the sub-code B97.7, which is papillomavirus as the cause of diseases classified in other chapters. We also found increased odds of dysplasia of cervix uteri (N87, OR=3.2 (2.9-3.5) and abnormal findings in specimens from female genital organs (R87, OR=2.3 (2.0-2.5). There is also an association with malignant neoplasm of cervix uteri (C53, OR=3.3 (1.9-5.7)) and other predominantly sexually transmitted diseases not elsewhere classified (A63, OR=1.3 (1.2-1.4)) - this also includes diagnosis code A63.0, which stands for anogenital (venereal) warts.

In total, 15 diagnosis codes with significantly different prevalences in cases and controls were identified (Figure 3.6).

3.3 Discussion

Our experimental work started with the annotation of GWAS analysis in which women who had at least one positive high-risk or potentially high-risk HPV test results were used as cases, and women who did not have the respective LOINC codes were used as controls, in total amounting to the final analysis including 3445 cases and 10467 controls. The GWAS analysis identified one locus for high-risk HPV infection, positioned in the MHC region on the short arm of chromosome 6 (6p21. 3).

The histocompatibility complex, known as the human leukocyte antigen (HLA) in humans, is the title given to a specific locus present in vertebrate DNA; this locus is characterized by several genes containing codes for producing specific cell surface proteins. The surface proteins these genetic codes produce play an essential role in boosting adaptive immunity; collectively, they are called MHC molecules [15]. The MHC molecules activate the adaptive immune system by linking to the peptide fragments taken from the invading pathogens; these peptide fragments are subsequently detected by the relevant T-cells, which, in most cases, destroy the pathogens. This immune mechanism triggers the activation of B-cells, and these B-cells produce antibodies, ultimately destroying the invading extracellular pathogenic particles; in the same way, this defense strategy brings macrophages into action, which engulf, digest, and destroy pathogens present in the intermolecular vesicles [15].

The signal that was detected in the MHC region on the short arm of chromosome 6 (6p21.3) alone did not tell us much about the specific HLA alleles that are associated with our trait of interest. We further mapped the signal to specific HLA alleles and found *HLA-DQB1**0603, *HLA-DRB1**1301, *HLA-DRB1**13, *HLA-DQA1**0103, *HLA-DQB1**0602 to be most significantly associated. The detected alleles were mentioned in previous GWAS association studies on HPV+ cervical cancer, particularly *HLA-DRB1**1301 and *DQB1**0603 alleles, were found to be associated with decreased risk of cervical cancer ([27], [29], [30]) and more broadly, the *HLA-DRB1**1301–*HLA-DQA1**0103–*HLA-DQB1**0603 haplotype conferred the strongest protection against cervical cancer [26]. *HLA-DQB1*0602* was found to be associated with high-risk HPV infection. Although these particular alleles were not detected in the genetic association studies of susceptibility to HPV infections, other alleles of the *HLA-DQB1* and *HLA-DQA1* genes have been identified, for example, in one case-control study of women with no HPV-related cancer who carried the alleles *DRB1*1503*, *DRB1*0395*, and *DQB1*0602* were found more likely to test positive for HPV [32], which is in line with our results. In a previous study, carriers of *DQB1*0602* and *DRB1*1501* were more frequent in the group with long-term HPV infections, indicating that these class II alleles contribute to the inability to clear an HPV infection [51]. While the previous genetic association studies of susceptibility to HPV infections had a small sample size and confined the studies to one or a few genes, ours is more trustworthy due to it being a GWAS association study and having a significantly bigger sample size of 3325 cases and 10153 controls.

The *HLA-DQB1* gene codes for a protein that plays an important function in the immune system. The *HLA-DQB1* gene is a member of the MHC class II gene family. MHC class II genes code for proteins that are found on the surfaces of some immune system cells [49]. Outside the cell, these proteins bind to protein fragments (peptides). These peptides are shown to the immune system via MHC class II proteins [49]. If the immune system detects the peptides as foreign (for example, viral or bacterial peptides), it launches an attack against the invading viruses or bacteria [49]. The protein encoded by the *HLA-DQB1* gene interacts with the protein encoded by another MHC class II gene, *HLA-DQA1* [49]. They combine to produce a functional protein complex called an antigen-binding DQ heterodimer [49]. This complex presents foreign peptides to the immune system to activate the body's immunological response [49].

Each MHC class II gene contains several potential variants, allowing the immune system to respond to a diverse spectrum of foreign invaders [49]. Researchers have discovered hundreds of distinct forms (alleles) of the *HLA-DQB1* gene, each of which is assigned a unique number (for example, *HLA-DQB1*06:02*) [49].

In the analysis of associated phenotypes, in total, 15 diagnosis codes with significantly different prevalences in cases and controls were identified. Women with a diagnosis of high-risk HPV infection have significantly more diagnoses of viral agents as the cause of diseases classified to other chapters, abnormal findings in specimens from female genital organs, malignant neoplasm of cervix uteri, and other predominantly sexually transmitted diseases, which includes anogenital (venereal) warts. These findings reflect what is already known about the HPV infection effects on one's health. Although there is no specific code for high-risk HPV infection, we see a strong association with the B97 code; in the future, we could try to run a GWAs for this code to see if we can find any novel associations.

We have faced several limitations during our research, including the absence of a specific code for high-risk HPV infection, which complicated the identification of individuals and restricted the sample size. Also, due to the absence of a replication cohort, everything we report should be replicated in independent datasets. In addition, with the vast majority of HPV infections being temporary and eliminated by the immune system within a few years, it is hard to be sure that controls have not had an HPV infection, in our case, a high-risk HPV infection.

In conclusion, there is still a lot of research that needs to be done concerning the genetics of high-risk HPV infection with a large sample size and a GWAS approach, preferably, so that multiple genetic variations across the genome of different individuals could be analyzed so that they are associated with high-risk HPV. Identifying more genetic susceptibility factors associated with high-risk HPV infection could improve prevention strategies and contribute to further advancements in personalized medicine. For example, if a woman tests positive for the alleles that are associated with an increased risk of HR-HPV infection, vaccination against HPV infection can be suggested by her healthcare provider.

SUMMARY

Cervical cancer is a prevalent and serious health concern among women worldwide, primarily caused by high-risk human papillomavirus (HPV) infection. This thesis aimed to investigate the genetic susceptibility factors associated with high-risk HPV infection using a genome-wide association study (GWAS) approach.

The GWAS analysis identified a significant genetic locus within the major histocompatibility complex (MHC) region on chromosome 6p21.3, associated with high-risk HPV infection.

HLA-DQB10603, *HLA-DRB11301*, *HLA-DRB113*, and *HLA-DQA10103* have been found to be associated with a decreased risk, and *HLA-DQB10602*, and *HLA-DRB1*15* with an increased risk of high-risk HPV infection These alleles have been previously linked to a reduced risk of cervical cancer or increased susceptibility to HPV infection in related studies, further supporting their relevance.

Additionally, we identified several diagnosis codes that exhibited significantly different prevalence in cases and controls, indicating potential associations with viral agents, abnormal findings in female genital specimens, malignant neoplasm of the cervix uteri, and other sexually transmitted diseases.

The findings contribute to the existing knowledge of the genetics of high-risk HPV infection. The results emphasize the importance of larger-scale GWAS studies to understand the genetic susceptibility factors underlying high-risk HPV infection. This understanding can enhance preventive strategies, aid in personalized medicine approaches, and ultimately contribute to global efforts to reduce the burden of cervical cancer. Further studies in diverse populations are needed to validate our results.

BIBLIOGRAPHY

[1] Mariann Koel, Urmo Võsa, Maarja Lepamets, Kristi Läll, Natàlia Pujol-Gualdo, Hannele Laivuori, Susanna Lemmelä, Mark Daly, Estonian Biobank Research Team, FinnGen, Priit Palta, Reedik Mägi, Triin Laisk. GWAS meta-analysis and gene expression data link reproductive tract development, immune response and cellular proliferation/apoptosis with cervical cancer and clarify overlap with other cervical phenotypes, 2021.

[2] Bruni L, Albero G, Serrano B, Mena M, Collado JJ, Gómez D, Muñoz J, Bosch FX, de Sanjosé S. ICO/IARC Information Centre on HPV and Cancer (HPV Information Centre). Human Papillomavirus and Related Diseases in the World, 2023.

[3] Natàlia Pujol Gualdo, Estonian Biobank Research Team, Reedik Mägi, Triin Laisk. Genome-wide association study meta-analysis supports association between MUC1 and ectopic pregnancy, 2022.

[4] Jo's Cervical Cancer Trust. The cervix. https://www.jostrust.org.uk/information/cervix/about-the-cervix

[5] Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L, Fischer K, Metspalu A. Cohort Profile. Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol, 2015.

[6] Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet [Internet]. 2016 Nov 1 [cited 2020 Oct 2];48(11):48–11.

[7] Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. 2017;25:25–27. Available from: https://pubmed.ncbi.nlm.nih.gov/28391899/.

[8] Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet [Internet]. 2018 Sep 1[cited 2020 Oct 2];50(9):50–9. Available from: https://pubmed.ncbi.nlm.nih.gov/30104761/OpenUrl

[9] Ahlbom A, Lichtenstein P, Malmström H, Feychting M, Pedersen NL, Hemminki K. Cancer in twins: genetic and nongenetic familial risk factors. Journal of the National Cancer Institute. 1997;89(4):287-293. https://doi.org/10.1093/jnci/89.4.287

[10] Bowden SJ, Bodinier B, Kalliala I, Zuber V, Vuckovic D, Doulgeraki T, Whitaker MD, Wielscher M, Cartwright R, Tsilidis KK, et al. Genetic variation in cervical preinvasive and invasive disease: A genome-wide association study. Lancet Oncol. 2021;22:548–557.

[11] Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Computational Biology. 2012;8(12):e1002822.

[12] Chen AA, Gheit T, Franceschi S, Tommasino M, Clifford GM. Human papillomavirus 18 genetic variation and cervical cancer risk worldwide. Journal of virology. 2015;89(20):10680-10687. https://doi.org/10.1128/JVI.01747-15

[13] Gray LA, Barnes ML, Lee JJ. Carcinoma-in-situ and dysplasia of the cervix. Annals of Surgery. 1960;151(6):951. https://doi.org/10.1097/00000658-196006000-00019

[14] Hirose Y, Onuki M, Tenjimbayashi Y, Mori S, Ishii Y, Takeuchi T, ... & Kukimoto I. Within-host variations of human papillomavirus reveal APOBEC signature mutagenesis in the viral genome. Journal of Virology. 2018;92(12):e00017-18.

[15] Janeway Jr CA, Travers P, Walport M, Shlomchik MJ. Principles of innate and adaptive immunity. In Immunobiology: The Immune System in Health and Disease. 5th edition. Garland Science. 2001.

[16] Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, ... & de Bakker PI. Imputing amino acid polymorphisms in human leukocyte antigens. PloS one. 2013;8(6):e64683. https://doi.org/10.1371/journal.pone.0064683

[17] Ramachandran D, Dörk T. Genomic risk factors for cervical cancer. Cancers. 2021;13(20):5137. https://doi.org/10.3390/cancers13205137

[18] Roy A, Matzuk MM. Reproductive tract function and dysfunction in women. Nature Reviews Endocrinology. 2011;7(9):517-525.

[19] Sankaranarayanan PW. Anatomy of the uterine cervix and the transformation zone - Colposcopy and treatment of cervical Precancer - NCBI bookshelf. National Center for Biotechnology Information. 2017. https://www.ncbi.nlm.nih.gov/books/NBK568392

[20] Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, ... & Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.

[21] Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nature Communications. 2017;8(1):1826.

[22] Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826. doi:10.1038/s41467-017-01261-5

[23] Ida-Tallinna Keskhaigla. Cervical cancer screening. (https://www.itk.ee/en/cervical-cancer-screening)

[24] Espinoza H, Ha KT, Pham TT, Espinoza JL. Genetic Predisposition to Persistent Human Papillomavirus-Infection and Virus-Induced Cancers. Microorganisms. 2021;9(10):2092. doi: 10.3390/microorganisms9102092. PMID: 34683414; PMCID: PMC8539927.

[25] Douglas JM Jr. In Goldman's Cecil Medicine (Twenty Fourth Edition), 2012.

[26] Chen D, Gyllensten U. A cis-eQTL of HLA-DRB1 and a frameshift mutation of MICA contribute to the pattern of association of HLA alleles with cervical cancer. Cancer Med. 2014;3(2):445-52. doi: 10.1002/cam4.192. Epub 2014 Feb 12. PMID: 24520070; PMCID: PMC3987094.

[27] Kamiza AB, Kamiza S, Mathew CG. HLA-DRB1 alleles and cervical cancer: A meta-analysis of 36 case-control studies. Cancer Epidemiol. 2020;67:101748. https://doi.org/10.1016/j.canep.2020.101748.

[28] Paaso A, Jaakola A, Syrjänen S, Louvanto K. From HPV Infection to Lesion Progression: The Role of HLA Alleles and Host Immunity. Acta Cytol. 2019;63(2):148-158. doi: 10.1159/000494985. Epub 2019 Feb 15. PMID: 30783048.

[29] Madeleine MM, Brumback B, Cushing-Haugen KL, Schwartz SM, Daling JR, Smith AG, et al. Human leukocyte antigen class II and cervical cancer risk: a population-based study. J Infect Dis. 2002;186(11):186–11. https://doi.org/10.1086/345980.

[30] Safaeian M, Johnson LG, Yu K, Wang SS, Gravitt PE, Hansen JA, et al. Human Leukocyte Antigen Class I and II Alleles and Cervical Adenocarcinoma. Front Oncol. 2014;4:119. https://doi.org/10.3389/fonc.2014.00119.

[31] Shrestha S, Wang C, Aissani B, Wilson CM, Tang J, Kaslow RA. Interleukin-10 gene (IL10) polymorphisms and human papillomavirus clearance among immunosuppressed adolescents. Cancer Epidemiol Biomarkers Prev. 2007;

[32] Maciag PC, Schlecht NF, Souza PS, Franco EL, Villa LL, Petzl-Erler ML. Major histocompatibility complex class II polymorphisms and risk of cervical cancer and human papillomavirus infection in Brazilian women. Cancer Epidemiol Biomarkers Prev. 2000;9:1183–1191.

[33] Bharadwaj M, Mehrotra R. In Animal Biotechnology, 2014

[34] Bernal-Silva S, Granados J, Gorodezky C, Aláez C, Flores-Aguilar H, Cerda-Flores RM, Guerrero-González G, Valdez-Chapa LD, Morales-Casas J, González-Guerrero JF, et al. HLA-DRB1 Class II antigen level alleles are associated with persistent HPV infection in Mexican women; a pilot study. Infect Agent Cancer. 2013;8:31. doi: 10.1186/1750-9378-8-31.

[35] Metcalfe S, Roger M, Faucher MC, Coutlée F, Franco EL, Brassard P. The association between human leukocyte antigen (HLA)-G polymorphisms and human papillomavirus (HPV) infection in Inuit women of northern Quebec. Hum Immunol. 2013;74:1610–1615. doi: 10.1016/j.humimm.2013.08.279.

[36] Ferguson R, Ramanakumar AV, Richardson H, Tellier PP, Coutlée F, Franco EL, Roger M. Human leukocyte antigen (HLA)-E and HLA-G polymorphisms in human papillomavirus infection susceptibility and persistence. Hum Immunol. 2011;72:337–341. doi: 10.1016/j.humimm.2011.01.010.

[37] Douglas JM Jr. In Goldman's Cecil Medicine (Twenty Fourth Edition), 2012.

[38] National Institutes of Health, National Cancer Institute. (24, May. 2008). https://www.cancer.gov/types/cervical/screening/abnormal-hpv-pap-test-results

[39] Shanmugasundaram S, You J. Targeting Persistent Human Papillomavirus Infection. Viruses. 2017;9:229. doi: 10.3390/v9080229.

[40] Ida-Tallinna Keskhaigla. Vaccination against HPV. Available from: https://www.itk.ee/en/patient/clinics/womens-clinic/vaccination-against-hpv.

[41] Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet [Internet]. 2017 Jun 1 [cited 2020 Nov 4];25(7):25–7. Available from: https://pubmed.ncbi.nlm.nih.gov/28391899/OpenUrl

[42] Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. PLoS One. 2013;8:e64683. Available from: https://dx.plos.org/10.1371/journal.pone.0064683.

[43] Kovalev G. Potential of Artificial Genomes in Genome-wide Association Studies, 2021.

[44] Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1826.

[45] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

[46] Del Río-Ospina L, Camargo M, Soto-De León SC, Sánchez R, Moreno-Pérez DA, Patarroyo ME, Patarroyo MA. Identifying the HLA DRB1-DQB1 molecules and predicting epitopes associated with high-risk HPV infection clearance and redetection. Sci Rep. 2020;10:7306. doi: 10.1038/s41598-020-64268-x.

[47] Adebamowo SN, Adeyemo AA, Consortium A.R.G.a.p.o.t.H.A. Classical HLA alleles are associated with prevalent and persistent cervical high-risk HPV infection in African women. Hum Immunol. 2019;80:723–730. doi: 10.1016/j.humimm.2019.04.011.

[48] Okunade KS. Human papillomavirus and cervical cancer. J Obstet Gynaecol. 2020;39:602-608. doi: 10.1080/01443615.2019.1633930.

[49] "HLA-DQB1." MedlinePlus, U.S. National Library of Medicine, National Institutes of Health, Genetics Home Reference, 10 May 2021. Available from: https://medlineplus.gov/genetics/gene/hla-dqb1/#resources.

[50] National Library of Medicine. (2022). rs3892710. Retrieved from https://www.ncbi.nlm.nih.gov/snp/rs3892710

[51] Beskow, A. H., Josefsson, A. M., & Gyllensten, U. B. (2001). HLA class II alleles associated with infection by HPV16 in cervical cancer in situ. International Journal of Cancer, Tumor Virology, 13 August 2001. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.1412

.

Appendix

R script for the calculating OR and CI for analysis of disease codes associated with a diagnosis of high-risk HPV

```
library(stringr)
library(dplyr)

# Read the original file
data <- read.table("~/Desktop/Thesis/phewas_hpv_for_phe-
was200Kmaincode.txt", header = TRUE)

# Remove rows containing S, T, U, V, W, X, Y subchapters
data <- data %>% filter(!str_detect(ICD, "S|T|U|V|W|X|Y"))

# Convert beta column to numeric, replacing non-numeric values
with NA
data$beta <- as.numeric(data$beta)

# Create a new data frame with selected columns
new_data <- data %>%
  mutate(OR = ifelse(is.na(beta), NA, exp(beta)),
         CI_U = ifelse(is.na(beta) | is.na(se), NA, exp(beta +
(1.96 * se))),
         CI_L = ifelse(is.na(beta) | is.na(se), NA, exp(beta -
(1.96 * se)))) %>%
  select(ICD, OR, CI_U, CI_L)

# Write the new data to a new file
write.table(new_data, "~/Desktop/Thesis/phewas_hpv_for_phe-
was200Kmaincode.new5.txt", sep = "\t", row.names = FALSE, quote =
FALSE)
```

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, Ksenia Chloe Bartlett

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

"Genetic Susceptibility Factors of High-Risk Human Papillomavirus (HPV)",

supervised by Triin Laisk and Mariann Koel.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ksenia Chloe Bartlett

24/05/2023