From DEPARTMENT OF MEDICAL EPIDEMIOLOGY AND BIOSTATISTICS
Karolinska Institutet, Stockholm, Sweden

GENETIC DETERMINANTS FOR SUSCEPTIBILITY, PROGRESSION AND PROGNOSIS OF PROSTATE CANCER

Robert Szulkin

Karolinska Institutet

Stockholm 2015

Published by Karolinska Institutet.
Printed by Eprint AB 2015

ISBN 978-91-7676-044-4

GENETIC DETERMINANTS FOR SUSCEPTIBILITY, PROGRESSION AND PROGNOSIS OF PROSTATE CANCER
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Robert Szulkin

Principal Supervisor:
Associate Professor Fredrik Wiklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor:
Associate Professor Mark Clements
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Professor Paul Franks
Lund University
Department of Clinical Sciences
Genetic and Molecular Epidemiology

Examination Board:
Associate Professor Fredrik Granath
Karolinska Institutet
Department of Medicine

Associate Professor Henrik Larsson
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Senior Clinical Lecturer Prabhakar Rajan
Queen Mary University of London
Barts Cancer Institute

Till Emma, Elliot och Alma

ABSTRACT

Prostate cancer is the most commonly diagnosed form of non-skin cancer among men in developed countries. Although a large proportion of patients eventually die from the disease, many indolent tumors are found via prostate specific antigen (PSA) testing. However, todays diagnostic tools are unable to distinguish small localized tumors that will have a benign development from early stage aggressive disease. Thus, over-diagnosis and over-treatment are two major concerns in prostate cancer management. Genetics have been shown to play an important role for prostate cancer initiation with an estimated heritability of 58% and over 100 identified single nucleotide polymorphisms (SNPs) associated with prostate cancer risk. However, much less is known about the involvement of genes in the progression and prognosis of the disease.

The overall objective of this thesis is to enhance the understanding of genetic determinants for initiation, progression and prognosis of prostate cancer. The purpose of **Study I** was to develop a prediction model for prostate cancer susceptibility, based on the current knowledge of genetic risk variants. Furthermore, we aimed to study the potential role of established prostate cancer risk variants in disease progression among men with a localized disease (**Study III**). In **Study II**, the heritability of prostate cancer-specific survival among diagnosed men was estimated and a genome-wide search for genetic determinants of the same outcome was performed in **Study IV**.

We found that a polygenic risk score model with 65 established prostate cancer risk SNPs and 68 novel variants optimally separates prostate cancer cases from healthy controls, with a prediction accuracy measured using the area under the curve (AUC) of 0.68. Furthermore, we observed that these 133 SNPs could be used for risk stratification; compared with an intermediate genetic risk score category (40%-60%), men with a low genetic risk score (lowest 5% percentile) had 84% decreased relative risk of prostate cancer and men with 5% highest risk scores had a four-fold increased relative risk.

Using a novel conditional likelihood approach for time-to-event data in brother pairs and father-son pairs, the heritability of prostate cancer survival was estimated to be 10%. We could also observe that common family environment had no effect (estimated to 0%) on prostate cancer survival. However, data simulations suggest that this may be underestimated. Furthermore, we could not find any association between SNPs and prostate cancer prognosis. None of 23 established prostate cancer risk SNPs investigated were found to be associated with disease progression in a cohort of men with localized disease. Moreover, in a genome-wide association study (GWAS) we did not find any association with prostate cancer survival at a genome-wide significant level.

In conclusion, with the current knowledge of prostate cancer genetics it is possible to identify men with high and low prostate cancer susceptibility risk. However, the predictive performance of established SNPs is not yet sufficient to be used alone in a screening program of prostate cancer. Furthermore, the findings in this thesis regarding prostate cancer

progression and survival suggest that development of prostate cancer and progression to lethal disease may be two separate biological mechanisms that involve different genes. In order to identify genetic risk variants associated with prostate cancer progression, future studies should be designed to find common variants with very low penetrance or rare variants with moderate to large effect.

LIST OF SCIENTIFIC PAPERS

I. Szulkin R, Whitington T, Eklund M, Aly M, Eeles RA, Easton D, Kote-Jarai Z, Amin Al Olama A, Benlloch S, Muir K, Giles GG, Southey MC, Fitzgerald LM, Henderson BE, Schumacher F, Haiman CA, Schleutker J, Wahlfors T, Tammela TL, Nordestgaard BG, Key TJ, Travis RC, Neal DE, Donovan JL, Hamdy FC, Pharoah P, Pashayan N, Khaw KT, Stanford JL, Thibodeau SN, McDonnell SK, Schaid DJ, Maier C, Vogel W, Luedeke M, Herkommer K, Kibel AS, Cybulski C, Lubiński J, Kluźniak W, Cannon-Albright L, Brenner H, Butterbach K, Stegmaier C, Park JY, Sellers T, Lim HY, Slavov C, Kaneva R, Mitev V, Batra J, Clements JA, BioResource, Spurdle A, Teixeira MR, Paulo P, Maia S, Pandha H, Michael A, Kierzek A, PRACTICAL Consortium, Gronberg H, Wiklund F.
Prediction of individual genetic risk to prostate cancer using a polygenic score.
Prostate. 2015 Sep; 75(13):1467-1474.

II. Szulkin R, Clements M, Magnusson P, Wiklund F, Kuja-Halkola R.
Estimating heritability of prostate cancer-specific survival using population-based registers.
Manuscript.

III. Szulkin R, Holmberg E, Stattin P, Xu J, Zheng S, Palmgren J, Grönberg H, Wiklund F.
Prostate cancer risk variants are not associated with disease progression.
Prostate. 2012 Jan;72(1):30-9.

IV. Szulkin R, Karlsson R, Whitington T, Aly M, Gronberg H, Eeles RA, Easton DF, Kote-Jarai Z, Amin Al Olama A, Benlloch S, Muir K, Giles GG, Southey MC, FitzGerald L, Henderson BE, Schumacher FR, Haiman CA, Sipeky C, Tammela TL, Nordestgaard BG, Key TJ, Travis RC, Neal D, Donovan JL, Hamdy FC, Pharoah PD, Pashayan N, Khaw KT, Stanford JL, Thibodeau SN, McDonnell SK, Schaid DJ, Maier C, Vogel W, Luedeke M, Herkommer K, Kibel AS, Cybulski C, Lubinski J, Kluzniak W, Cannon-Albright L, Brenner H, Herrmann V, Holleczek B, Park JY, Sellers TA, Lin HY, Slavov C, Kaneva RP, Mitev VI, Batra J, Clements JA, Spurdle A, Teixeira MR, Paulo P, Maia S, Pandha HS, Michael A, Kierzek A, Albanes D, Andriole GL, Berndt SI, Chanock SJ, Gapstur SM, Giovannucci EL, Hunter DJ, Kraft P, Le Marchand L, Ma J, Mondul AM, Penney KL, Stampfer M, Stevens VL, Weinstein SJ, Trichopoulou A, Bueno-de-Mesquita HB, Tjonneland A, Cox DG, Maehle L, Schleutker J, Lindstrom S, Wiklund F.
Genome-wide association study of prostate cancer-specific survival.
Cancer Epidemiol Biomarkers Prev. 2015 Aug 25 [Epub ahead of print].

CONTENTS

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic Acid
DRE	Digital Rectal Examination
GWAS	Genome-wide Association Study
HWE	Hardy Weinberg Equilibrium
LD	Linkage Disequilibrium
PIN	Prostatic Intraepithelial Neoplasia
PRACTICAL	Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (Prostate cancer genetics consortium)
PSA	Prostate Specific Antigen
QC	Quality Control
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism

1 INTRODUCTION

Prostate cancer is a major health concern over the whole world (**Figure 1**), particularly in more developed countries (including North America, Australia, New Zealand, and Northern/Western parts of Europe). It was the most common non-skin cancer among men in the developed parts of the world with an estimated incidence of 759,000 per year and 142,000 deaths (3rd leading cause of cancer death) in year 2012[1]. Since prostate cancer is most common among older men and the population is ageing, prostate cancer incidence is expected to increase over the coming years. In the absence of potentially modifiable risk factors, primary prevention is difficult so that early detection and treatment will become increasingly important. Many of the tumors that are found today are small, localized and so slowly growing that they will not cause any symptoms to the diagnosed man. A major problem is that today's diagnostic and prognostic tools are insufficient to separate these indolent tumors from early stage aggressive disease. This has resulted in a considerable over-treatment and over-diagnosis of the disease. Thus, the identification of biomarkers that can improve diagnosis and prediction of prognosis of prostate cancer is warranted.

In this thesis we have investigated whether genetics can be used for this purpose. In **Study I** we have assessed how well we can predict the risk of developing prostate cancer, based on the current knowledge in genetics, which could be of importance in a screening situation. In studies **II-IV** we have studied the role of genetics in the prognosis (progression and survival) of the disease.

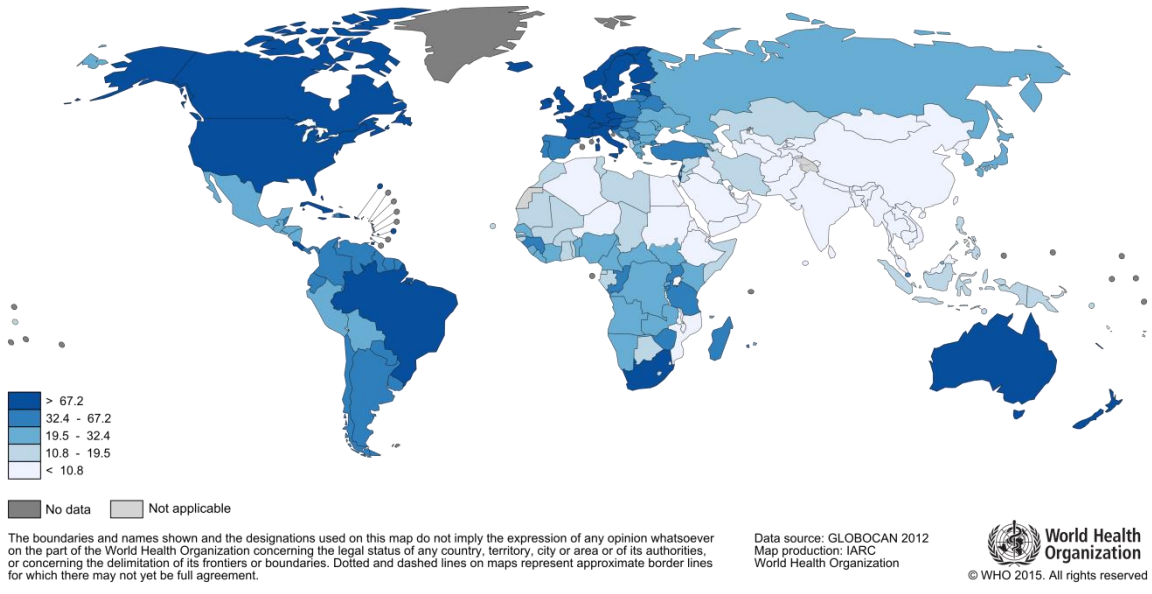

Figure 1: *Age standardized prostate cancer incidence rates per 100,000 individuals (standardized to Segi's World Standard Population[2]).*

2 BACKGROUND

2.1 PROSTATE CANCER

2.1.1 The prostate

The prostate is a gland organ that is located in the male pelvis, just below the urinary bladder in front of the rectum, surrounding the urethra (**Figure 2**). A healthy adult prostate has the size of a walnut. The organ has three anatomical zones: (i) the peripheral zone which is closest to the rectum, (ii) the transition zone lies adjacent to the urethra (surrounding it) and (iii) the central zone located between the peripheral and transition zones[3]. In a young man, each of these zones constitutes approximately 65%, 25% and 10% of a normal prostate[4]. All three anatomical zones contain epithelial cells that produce organ specific enzymes, prostatic acid phosphate (PAP) and prostate specific antigen (PSA). The prostate plays an important role in the male reproductive system. During ejaculation, seminal fluids from the seminal vesicles are mixed with prostatic secretion. PSA facilitates the movement of the sperm to fertilize the ovulated egg, by liquefying coagulated semen[5].

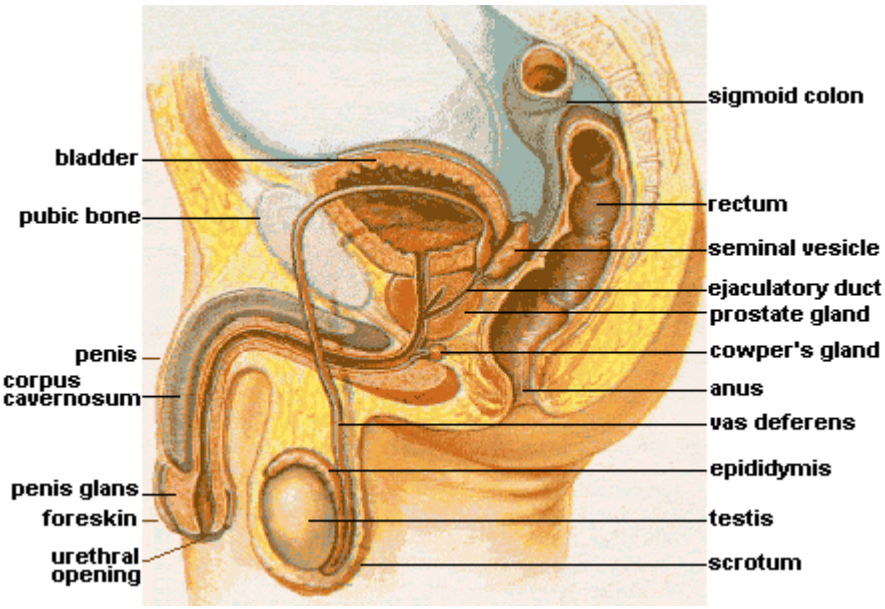

Figure 2: *Male internal sexual anatomy. Reprinted from Wikimedia Commons[6] under the license of Creative Commons.*

2.1.2 Diseases of the prostate

As men grow older various diseases of the prostate gland become more common. *Benign prostatic hyperplasia* (*BPH*) emerges from a natural growth of the transition zone (surrounding the urethra) of the prostate, which can begin from as early as age 30 years[7]. BPH is a chronic condition that is progressive with age[8], and results in difficulties in urination for approximately 40% of the male population before age 70 years[9,10]. Since BPH and

prostate cancer usually arise from different zones of the prostate (although the diseases can coexist in the same region)[4,11,12], BPH is in general not viewed as a cause of prostate cancer. However, there is emerging evidence for a possible link between BPH and prostate cancer via chronic inflammation[8,13]. Prostatitis is an inflammation (usually non-bacterial[14]) of the prostate, which is also common among men aged 30 years or more[15]. Approximately 50% of all men experience an episode of prostatitis sometime during their lifetime[16]. A recent hypothesis is that inflammation is a key event in prostate cancer development, although the etiology remains unclear[17-22].

2.1.3 Prostate cancer

Cancer is the most severe disease that can occur in the prostate. A large American autopsy study of men who died from other causes than prostate cancer have shown that approximately 45% in the age group 50-59 and approximately 70% in the age group 70-79 have a latent prostate cancer (tumor cells in the prostate which would have been found in a needle biopsy)[23,24]. Cancers in the prostate are almost exclusively *adenocarcinoma*, defined as tumors that originate from malignant growth in epithelial tissues. Approximately 70% of the tumors arise in the peripheral zone, 25% in the transition zone, 5% in the central zone[4], and often tumors invade several zones. Furthermore, prostate cancer is commonly multifocal, where the prostate harbours multiple tumors[25]. Lately, evidence has emerged which supports a theory that these develop independently and simultaneously from lesions of Prostatic Intraepithelial Neoplasia (PIN[26]; sometimes acknowledged as a histopathological precursor state of prostate cancer)[27,28].

Advancement of the disease often involves a tumor perforating the prostate capsule to nearby tissues in the pelvis, urinary bladder, urethra and seminal vesicle. Metastatic spread occurs primarily via lymph nodes in the pelvis and thereafter to the bones. In rare cases the prostate cancer also spreads to the lungs and other organs[29]. Some tumors are more aggressive than others, such that they spread more quickly outside the prostate gland. Although tumors that arise in the central zone are rare, they seem to be most prone to metastasis[30], while cancers in the peripheral zone are of intermediate metastatic potential[31-33]. Furthermore, a current hypothesis is that the metastatic spread originates from one single clone, that is, from only one of the independent tumors within the prostate [34,35]. Thus, given that prostate cancer may consist of several independent tumors that can evolve to become aggressive, the diagnosis and prognostic prediction of prostate cancer is not straightforward.

2.1.4 Diagnosis and prognosis

Today, prostate cancer is most commonly detected via a PSA test, and sometimes by clinical symptoms (for example, difficulties to empty the bladder). In the early stages of the disease, when the tumor is confined to the prostate, symptoms are rare. Metastasized disease often appears as skeletal pain in the back, hips or pelvis[36,37]. A first step in a prostate cancer diagnosis typically involves analysis of the PSA test and a palpation of the prostate, also called a digital rectal exam (DRE). Various guidelines for further evaluation exist[36]. Swedish

national guidelines regarding PSA levels are age dependent to limit over-diagnosis of clinically insignificant cancer. Given a negative DRE, the following thresholds are recommended for further investigation of prostate cancer[38]:

- PSA≥2µg/l for men aged<50 years.
- PSA≥3µg/l for men aged 50-70 years.
- PSA≥5µg/l for men aged 70-80 years.
- PSA≥7µg/l for men aged>80 years.

A follow-up usually involves a transrectal ultrasound guided needle biopsy, where 12 cores are sampled from the prostate, followed by a histopathological evaluation of the specimens. Results are reported by using the Gleason grading system with scores that range from 1 (for well differentiated prostate glandular cells) to 5 (for poorly differentiated glandular cells)[39]. The sum of the most prevalent Gleason pattern and the highest graded area is reported and is then used as an indicator of tumor aggressiveness. Before year 2005 the sum of the two most prevalent Gleason patterns was used[40].

Radiology is used to establish the stage of the tumor in the TNM classification system. The following main classification categories are defined by the American Joint Committee on Cancer (AJCC)[41,42]:

T stage (Primary tumor)
TX: Primary tumor cannot be assessed.
T0: No evidence of primary tumor.
T1: Clinically inapparent tumor neither palpable nor visible by imaging.
T2: Tumor confined within prostate.
T3: Tumor extends through the prostate capsule.
T4: Tumor is fixed or invades adjacent structures other than seminal vesicles, such as external sphincter, rectum, bladder, levator muscles, and/or pelvic wall.
Some sub-categories exist, for example, T1c are tumors identified by needle biopsy (e.g. because of elevated PSA).

N stage (Regional lymph nodes)
NX: Regional lymph nodes were not assessed.
N0: No regional lymph node metastasis.
N1: Metastasis in regional lymph node(s).

M stage (Distant metastasis)
M0: No distant metastasis.
M1: Distant metastasis.

Localized prostate cancer is usually characterized by a slow growth of the tumor. However, it is difficult to distinguish an indolent disease (that never leads to any symptoms or death) from an early stage tumor that will develop aggressively[43]. Today's diagnostic tools (PSA, Gleason

and TNM staging) perform sub-optimally in this task. The following prognostic risk groups are commonly defined[36]:

- Low risk: T1-T2a, Gleason sum ≤6 and PSA<10 µg/l.
- Intermediate risk: T2b or Gleason sum 7 or PSA 10-20 µg/l.
- High risk: T2c-T3 or Gleason sum 8-10 or PSA>20 µg/l.

2.1.5 Prostate cancer screening

Screening for prostate cancer using PSA has been controversial due to the shortcomings of the PSA test. The main issue with the PSA test is that it is difficult to find a cut-off value that yields both a good sensitivity and specificity. For example, a Swedish study reported that cut-off values at 1, 3, 4, and 5 µg/l resulted in sensitivities of 96%, 59%, 44% and 33%, and specificities 44%, 87%, 92%, and 95%, other studies show similar results[44-46]. Two large PSA screening trials have evaluated the effect of PSA screening: the American Prostate, Lung, Colorectal, Ovarian (PLCO) cancer screening trial and the European Randomized Study of Screening for Prostate Cancer (ERSPC) study. The PLCO study did not find any mortality reduction for those who attended a PSA screening program[47]. However, this study has been criticized for a high proportion of opportunistic screening in the control arm participants and poor biopsy compliance. The level of opportunistic screening was less of an issue in the ERSPC study because PSA testing was introduced later in the participating European countries. The European trial reported a 21% relative risk reduction of prostate cancer mortality (38% in the Gothenburg sub-cohort of ERSPC) for those who were screened after 13 years of follow-up[48]. However, this came with a cost of considerable over-diagnosis and over-treatment. It was estimated that 781 men needed to be screened and 27 to be diagnosed to prevent one death from prostate cancer. As a result of this, no national prostate cancer screening programs exist. Nevertheless, opportunistic screening is very common[49] and there is an urgent need for better biomarkers to prevent over-diagnosis and over-treatment.

2.1.6 Treatment

Active surveillance is usually recommended to patients with a localized low risk tumor. Patients are initially not treated but monitored (repeated PSA testing and DRE) until the occurrence of either progression of disease or a change in preference for treatment. Furthermore, patients with limited life expectancy, such as due to advanced age or substantial comorbidities, are also less likely to receive curative treatment. Watchful waiting is an alternative strategy where the disease is monitored (without treatment) until clinical symptoms appear, followed by hormonal treatment (Androgen deprivation therapy; ADT). This is recommended to patients with localized disease without clinical symptoms and with a limited life expectancy.

Patients with an intermediate risk disease (as defined above) benefit from curative treatment, including surgical removal of the prostate (prostatectomy) or radiation therapy. These treatments may cause temporary or persistent adverse effects, including incontinence and

impotence. Cancers that progress to locally advanced or metastatic disease are usually not treated with curative intent but to prevent the tumor from further advancement. The patient is usually treated with hormones (ADT) to reduce testosterone levels. However, eventually most advanced prostate cancers progress and become castration resistant (i.e. the tumor is no longer dependent of testosterone). At this stage the patients with metastatic disease are offered palliative treatment[29,36,37].

2.1.7 Incidence, prevalence and mortality

In **Figure 3** we can see that prostate cancer incidence has increased steadily since year 1960 in USA, Sweden and UK. The introduction of PSA testing in the begging of the 1990s has resulted in a rapid growth of the incidence. This increase is mostly due to PSA detected small, localized tumors that have excellent prognosis. Thus, mortality rates have been relatively stable with a somewhat decreasing trend. As a result of this, the 10-year relative prostate cancer survival in Sweden has almost doubled, from 44% among men diagnosed between 1989 and 1993 to 79% among men diagnosed between 2009 and 2013[50]. Furthermore, this has led to an increased prevalence of the disease (1922.7 per 100,000 Swedish men lived with prostate cancer in 2013).

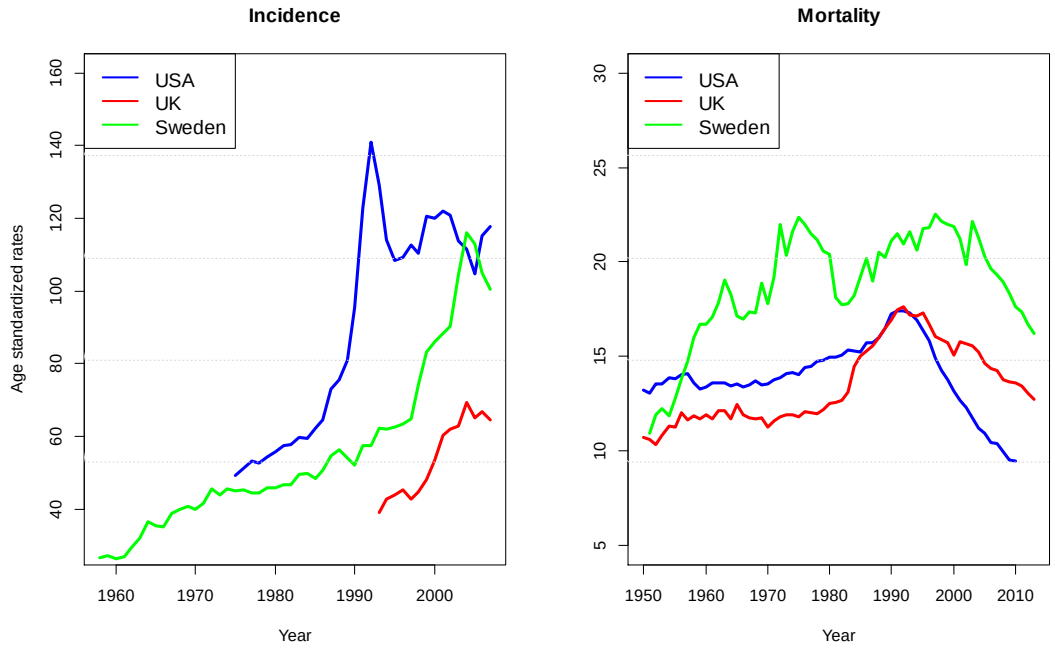

Figure 3: *Age standardized incidence and mortality rates per 100,000 men in USA, Sweden and UK[51,52] (standardized to Segi's World Standard Population[2]).*

2.1.8 Risk factors

Prostate cancer is suggested to arise from a combination of genetic, lifestyle and environmental factors. Apart from age, family history of prostate cancer and ethnicity are the only established risk factors for the disease. Men with a diagnosed first-degree relative have a

2-3 fold increased risk of getting prostate cancer compared to men without a family history[53]. The familial risk increases with the number of diagnosed family members and with age at onset of the relatives[54]. The heritability of prostate cancer susceptibility was recently estimated to 58% in a Nordic twin study[55], suggesting that the major part of the familial aggregation is due to genetic factors. Furthermore, prostate cancer survival has also shown concordance in families. The risk of dying from prostate cancer is increased if the father had poor prostate cancer survival[56-58]. However, the heritability of prostate cancer survival has never been reported.

The highest incidence rates are found among men with African American ancestry (in particular among those living in USA)[59,60], followed by Scandinavians. Asians have almost ten-fold lower incidence than white men living in developed countries[61]. However, the incidence increases significantly among Asian men who moved to western countries and adapted the local lifestyle[62,63]. This suggests that environmental factors may have a substantial role in prostate cancer incidence. Numerous life style factors have been associated with an increased risk of prostate cancer incidence (including dairy products, meat cooked at high temperature, saturated fatty acids and sexually transmitted diseases) and other have been reported to have a protective effect (tomatoes, green tea and soy products)[21,64,65]. However, these findings are inconsistent and not well understood. Furthermore, few established environmental or life style risk factors for prognosis or prostate cancer survival exist[66], with the possible exception of BMI[67-69].

2.2 PROSTATE CANCER GENOMICS

2.2.1 The human genome

Genetic information is transferred from parents to children via a molecule called Deoxyribonucleic acid (DNA). All cells in the body, except the red blood cells, harbor DNA in the nucleus. The DNA molecule has two very long complementary strands which form a so called *double helix* structure (**Figure 4**). Each strand has a deoxyribose sugar-phosphate backbone, to which nucleotide bases are attached. There are four different bases: *adenine (A), thymine (T), cytosine (C)* and *guanine (G).* Nucleotides on one strand are complementary to nucleotides on the other strand; A always binds to T and C to G via hydrogen bonds. The human DNA is composed of 23 chromosomes, 22 autosomal and 1 sex chromosome (X or Y). Most cells in the body have two copies of each chromosome (diploid), one from the mother and one from the father. Only germ cells (sperm and egg) have one copy of each chromosome (haploid). These are formed in a process called *meiosis*, where genetic material is exchanged (*meiotic recombination*) between the two chromosome copies[70].

A fundamental function of DNA is to code for proteins, which carries out the function of a cell. The first step of this process is that complementary molecules of messenger ribonucleic acids (mRNA) bind to one of the separated DNA strands, with the help of the enzyme RNA polymerase. The mRNA molecule is a single stranded molecule with the same nucleotide bases as DNA, except for thymine (T) which is replaced by *uracil* (U). The genetic

information is transported by mRNA from the nucleus to the cytoplasm of the cell, where it is translated to a protein by ribosomes and transfer RNA (tRNA). Triplets of RNA nucleotides are translated to amino acids and finally synthesized to proteins[70].

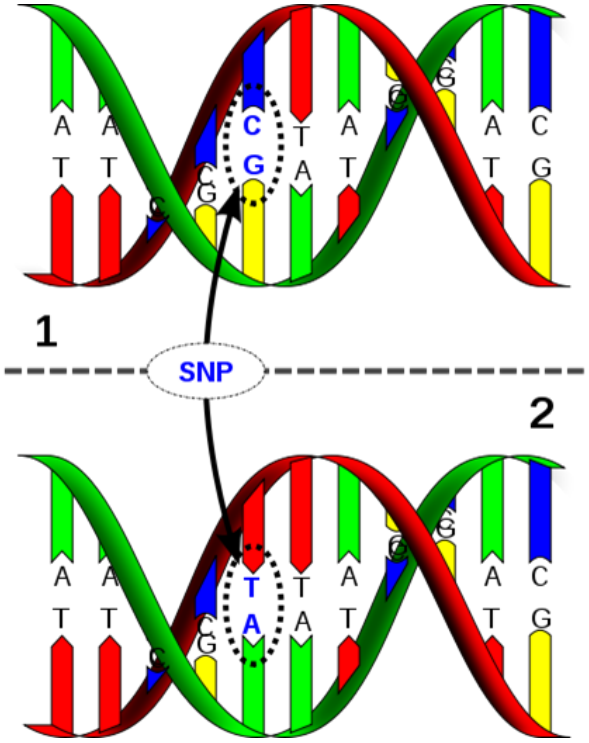

Figure 4: *Two aligned DNA molecules, which differs in one base-pair (**SNP**). Reprinted from Wikimedia Commons under the GNU free Documentation Licence [71].*

Genes are discrete regions of DNA which encodes for proteins. The initiation of the transcription of a gene is regulated by a part of the gene that is called the *promoter*. A gene consists of overlapping protein coding (*exons*) and non-coding (*introns*) sequences of DNA. The human genome consists of approximately 20,000 protein coding genes[72,73], a very small part of the genome (approximately 1%)[74]. However, several other important functional elements exist, for example regulatory elements (promoters, enhancers, silencers), DNA methylation sites, RNA transcribed regions , transcription-factor binding sites and open chromatin structures (DNaseI hypersensitive sites)[73,75]. Lately, the understanding of the function of these non-coding regions has increased, to large extent, via the Encyclopedia of DNA Elements (ENCODE) project, which has listed numerous functional elements for various cell types (including for example the prostate cancer lymph-node metastases cell line called LNCaP)[75].

2.2.2 Single nucleotide polymorphism (SNP)

The human genome has approximately 3 billion nucleotide base-pairs, 99% of these are the same in all humans. Positions (*loci;* singular: *locus*) where mutations have occurred and spread to more than 1% of the population during evolution are called polymorphic. There are

various kinds of variations (polymorphisms), for example, copy number variations (CNVs) and indels, which both are insertions and deletions of DNA. However, the most common polymorphisms are single nucleotide polymorphisms (SNPs). These are point mutations, where one nucleotide has been replaced by another (**Figure 4**), A↔C or C↔T substitutions (in both directions) are most frequent[70]. Thus, a SNP can only have two possible variants (*alleles*). Individuals with two copies of the same allele are called *homozygous*, otherwise *heterozygous* (if the two alleles are different). SNPs are the genetic markers that have been studied in this thesis.

To this date, millions of SNPs have been identified, much as a result of large genetic sequencing studies as the international HAPMAP project[76], the UK10K project[77] and the 1000Genome project[78]. The latter is an international collaboration, which has the aims to identify human genetic variations of all types and to define haplotype structures (sequences of SNPs on the same chromosome). We used these data for imputation, as described in section **5.4.4**. SNPs that are located inside of exons can be either *synonymous* or *nonsynonymous*, depending on whether they change the amino acid sequence or not. The latter category is further divided into *missense* (changes one amino acid) or *nonsense* (introduces a stop sequence which truncates the protein) mutations. In general, the majority of all SNPs in the genome are situated in introns and intergenic regions (so called "gene deserts"). Thus, they do not have a direct protein coding function.

As discussed in section **2.1.8** genetic factors are major components of prostate cancer development. A graph of all currently known prostate cancer risk variants is shown in **Figure 5**. Three rare variants (minor allele frequency (MAF) <1%) with moderate effect sizes have been identified in candidate gene studies: two breast cancer predisposition genes (BRCA1 and 2) and HOXB13. To this date, three germline mutations in the tumor suppressor BRCA2 gene are associated with the highest prostate cancer risk (8.6-fold for young onset disease)[79,80]. Furthermore, a mutation in BRCA2 has been associated with poorer survival[81,82]**.** Carriers of mutations in BRCA1 or in HOXB13 have been associated with an approximately 3.5-fold increased risk of disease susceptibility[83,84]. The HOXB13 variant is likely a Nordic founder mutation and is more frequent in these countries[84]. A functional study of HOXB13 has implicated that the gene is associated with disease prognosis (tumor progression and metastasis), however not with prostate cancer survival[85]. In total, the three rare variants explain a very small proportion of the familial risk of prostate cancer.

To date, approximately 100 common SNPs (MAF>1%) with low to moderate effect sizes have been identified in Genome-wide association studies (GWAS)[86-93]. Most of them were found via the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium, which will be given a more detailed description in **section 4.2**. In total, SNPs in prostate cancer GWAS regions, explain approximately 39% of the familial risk of prostate cancer[94]. However, since the effect sizes of these SNPs are low (most with relative risks<1.5) or most often very low (1.04-1.10) their alone predictive capacity of prostate cancer incidence is limited. Nevertheless, the cumulative

effects of SNPs have proven to be more relevant[88,95]. In **study I** of this thesis we constructed a prediction model for prostate cancer, based on current knowledge of prostate cancer genetic risk variants.

Most of the known prostate cancer risks SNPs, identified via GWAS, are located in introns or gene deserts. Several fine-mapping studies that aim to identify the functional variant and explore its functional role have emerged lately[94,96,97]. For example, in a recent fine-mapping study of 64 known prostate cancer regions, performed by PRACTICAL, an expression quantitative trait loci (eQTL) analysis revealed that 20% of the regions were associated with gene expressions in prostate tumor tissue[94]. Furthermore, several known prostate cancer risk regions have shown functional associations with genes[87,94,97-99]. However, these findings needs to be further explored in future studies.

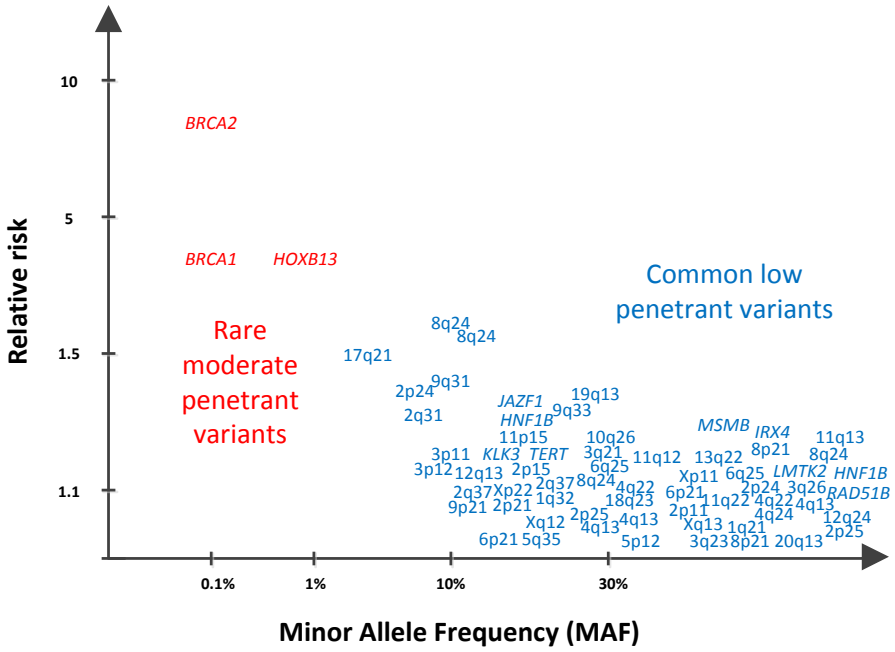

Figure 5: *Overview of prostate cancer risk variants. Figure was adapted from Goh and coworkers*[100].

In general, the established prostate cancer risk variants found in GWAS do not distinguish aggressive from less aggressive disease. Some of the SNPs have been reported to discriminate between aggressive cases and healthy controls but in case-only analysis (which is clinically more relevant) they were not associated with prognostic outcomes, such as prostate cancer survival[86,88,101]. One exception is the established risk SNP rs2735839, located near the *KLK3* gene at chromosome 19q13, that has been found associated with prostate cancer survival[102]. However, the reported prostate cancer risk allele was inversely associated with disease prognosis[102,103]. Furthermore, Lin and coworkers identified SNPs that were associated with the trait in a candidate gene study[104] and the results were replicated in independent studies[105,106]. Moreover, recently two common variants on chromosome 3q26 and 5q14 were found to be associated with Gleason score in a case-only GWAS[107]. However, currently no genetic variants have been associated with prostate cancer survival at a genome-wide significant level[108]. In **study IV** of this thesis we performed the largest prostate cancer survival GWAS to this date.

3 AIMS

The overall objective is to advance the understanding of genetic determinants for initiation, progression and prognosis of prostate cancer. We have the following specific aims, which correspond to each the four studies in this thesis:

I. To predict prostate cancer susceptibility using a polygenic risk score.
II. To estimate the heritability of prostate cancer survival.
III. To assess association between established prostate cancer susceptibility variants and disease progression.
IV. To perform a genome-wide assessment of genetic determinants for prostate cancer survival.

4 STUDY POPULATIONS

4.1 GENOME-WIDE ASSOCIATION STUDY (GWAS) POPULATIONS

4.1.1 CAPS GWAS

CAPS (CAncer of the Prostate in Sweden) is a Swedish population-based case-control study of prostate cancer[95,109,110]. Cases were identified from four of the six regional cancer centers in Sweden; western (Gothenburg region) and southern regions (Malmö area) were not included. All men from the northern and central parts of Sweden aged less than 80 years and from the Stockholm area and southeastern part of Sweden less than 65 years were asked to participate if they had a newly diagnosed (biopsy or cytologically confirmed) adenocarcinoma of the prostate. Recruitment was ongoing in two stages from July 2001 to October 2003. Virtually complete coverage was achieved since it is compulsory by Swedish law to report all cancers to the regional cancer centers. Out of 3,648 identified men 3,161 (87%) agreed to participate in the study. Through record linkage to the National Prostate Cancer Register (NPCR) (described in **section 4.5**), clinical information regarding diagnostic PSA-levels, Gleason score and TNM stage at diagnosis was obtained.

A random selection of controls from the Swedish Population Register was concurrently recruited with the cases. Controls were frequency matched to cases on age (in 5-year intervals) and region of residence. In total 2,149 (68% of all invited) control subjects agreed to participate in the study. All participants (both cases and controls) in the study replied to a questionnaire regarding risk factors for prostate cancer (including family history) and donated a blood sample at enrollment. Vital status of all study participants and cause of death among deceased participants is annually assessed via record linkage to the Swedish Cause of Death Register.

Genotyping for GWAS was performed in two phases - CAPS1 and CAPS2. In the first stage, 498 cases with aggressive prostate cancer (patients that met any of the following criteria at diagnosis: Gleason score ≥8, diagnostic PSA >50 ng/mL, T3/4 stage or present metastases) and 494 matched controls were included[111]. Subsequently an additional 1,475 cases and 527 controls (637 with aggressive disease and 838 with less aggressive Gleason 6 disease) were genotyped in the second stage. The genotyping was performed by collaborators at the Wake Forest University, USA. Two sets of arrays were used, GeneChip Human Mapping 500K (CAPS1) and 5.0K (CAPS2) from Affymetrix (Santa Clara, CA)[86]. In **Study IV**, we followed 1,985 cases (from CAPS1 and CAPS2) for prostate cancer survival, of which 545 had died from the disease before the end of follow-up in December 31, 2012.

4.1.2 UKGPCS1 GWAS

Prostate cancer cases from a large nationwide United Kingdom study, established in 1993, called UK Genetic Prostate Cancer Study[87,90,112] (UKGPCS), were ascertained to perform a GWAS. Cases were eligible if they had a clinically detected tumor (i.e. not PSA screening detected without any symptoms of prostate cancer) and were less than or equal to 60 years at

diagnosis or had a strong family history of prostate cancer. Controls were recruited from a national PSA screening study, ProtecT, if they were above the age of 50 and had a low PSA value (<0.5 ng/ml). In total, 1,906 prostate cancer cases and 1,934 controls were selected for this first GWAS stage. Genotyping was performed on an Illumina Infinium HumanHap550 array, generating 534,446 SNPs. In **Study IV**, we followed 1,783 cases for prostate cancer survival (457 cases died from the disease) up to the end of year 2012. Cause of death is ascertained every third month by linkage to national registers.

4.1.3 UKGPCS2 GWAS

The top findings in the UKGPCS1 GWAS (SNPs with $P<10^{-6}$) were assessed for replication in a subsequent case-control study, called UKGPCS2. In total 47,120 SNPs were genotyped on Illumina iSELECT assays. Participants in this study, 3,268 cases and 3,366 controls were recruited from UK and Australia (Melbourne area)[89,90]. In **study IV** of this thesis, 772 cases from UK (same inclusion criteria as in UKGPCS1) with available follow-up data for prostate cancer survival were included, of which 189 patients had died from prostate cancer.

4.1.4 BPC3 GWAS

Seven studies from the BPC3 consortium **(**Breast and Prostate Cancer Cohort Consortium) were used for a GWAS of aggressive prostate cancer[113,114]. All included patients had a tumor with a Gleason score ≥8 or stage C or D (approximately equivalent to T3 and T4). In total 2,782 cases were followed for prostate cancer mortality out of which 598 died from the disease. Participants were genotyped on Illumina 610 or 610K SNP arrays. In **study IV** of this thesis, we assessed summary results (not individual data) from the BPC3 consortium.

4.2 PRACTICAL

Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) is a large prostate cancer genetics consortium, which initially was formed to find genetic variants associated with prostate cancer risk. Currently, over 86 research groups from all over the world are a part of this collaboration with over 200,000 samples available for genotyping[115].

PRACTICAL is part of an even larger partnership, Collaborative Oncological Gene-environment Study (COGS) together with three other cancer genetics consortia (breast, ovarian and *BRCA1/2* mutation carriers), which had the aim to create a custom made SNP chip (iCOGS) relevant for these cancers. This resulted in the production of an Illumina Custom Infimum SNP array with totally 211,155 SNPs. The collaboration within COGS has recently evolved (now also including lung and colorectal cancer) with the purpose to design a new SNP chip with approximately 600K SNPs, called the OncoArray. The manufacturing process of this chip and genotyping of approximately 110,000 samples from the PRACTICAL consortium has recently finished. Initial results from OncoArray genotyping are expected before the end of 2015. Our research group nominated approximately 3,000 markers from a top list of SNPs associated with prostate cancer survival to the OncoArray.

Out of the 211,155 SNPs on the iCOGS chip, 85,278 SNPs were suggested by PRACTICAL to be relevant for prostate cancer. These variants were used to develop a prediction model for prostate cancer in **Study I**. The main part (74,001) of these SNPs was chosen on the basis of a meta-analysis of the above described GWAS studies as well as the CGEMS[116] study (a publically available NIH funded GWAS). SNPs that showed evidence of association with prostate cancer risk, aggressive prostate cancer (as measured by the Gleason score), cause-specific mortality and early onset disease (age <55 at diagnosis) were selected for the chip design. Moreover, 13,739 SNPs were chosen from fine-mapping of 27 previously known prostate cancer susceptibility regions and 1,398 markers from candidate genes (such as hormone metabolism, HOX class of genes, the cell cycle and DNA repair)[88,117].

Table 1: *Age distribution among cases and controls in PRACTICAL.*

	Cases		Controls	
Study	**N**	**Age, mean(sd)**	**N**	**Age, mean(sd)**
CAPS	1,153	66.1 (7.8)	664	67.8 (7.5)
CPCS1	849	69.4 (7.9)	2,771	56.3 (15.3)
CPCS2	265	64.5 (6.4)	1,015	55.6 (15.3)
EPIC	722	64.9 (5.6)	1,079	59.6 (6.5)
EPIC-Norfolk	484	72.1 (7.6)	917	73.6 (9.6)
ESTHER	313	65.5 (5.1)	318	65.4 (5.3)
FHCRC	838	59.8 (7.2)	784	59.4 (7.2)
MAYO	767	65.1 (6.4)	488	65.5 (9)
MCCS	1,685	57.8 (8.4)	1,183	52.1 (8.6)
MEC	819	69.5 (7.6)	829	70.5 (8.5)
MOFFITT	455	64.7 (8.3)	130	61.5 (10)
PCMUS	151	68.8 (8.9)	140	66.9 (8.3)
Poland	438	67.5 (8.1)	359	62.8 (10.3)
ProtecT	1,563	62.8 (5.1)	1,474	59.6 (5.4)
QLD	186	61.4 (6.9)	87	69.2 (5.6)
SEARCH	1,371	63.1 (4.8)	1,244	54.4 (7.4)
STHM1	2,006	66.2 (7)	2,224	67.1 (6.7)
TAMPERE	2,754	68.2 (8)	2,413	NA
UKGPCS	4,549	63.8 (8)	4,182	58.2 (5.3)
ULM	603	63.8 (6.7)	354	58.4 (11.8)
UTAH	440	62.5 (8.9)	245	66.7 (9.7)
Total	**22,411**	**64.8 (8.0)**	**22,990**	**60.6 (10.7)**

In total, samples from 25,074 prostate cancer cases and 24,272 controls from 32 studies in PRACTICAL were genotyped on the iCOGS-chip. A subset of the studies from PRACTICAL that were included in **Study I** of this thesis are summarized in **Table 1** (distribution of cases and controls, together with their age) and **Table 2** (study design). Study populations with available follow-up data on prostate-cancer survival were included in **Study IV** (Epic-Norfolk, MOFFIT, Poland, ProtecT and QLD were excluded). The majority

Table 2: *Study designs in PRACTICAL.*

Study acronym	Study Name	Country	Design
CAPS	Cancer of the Prostate in Sweden	Sweden	Case-control, Population-based[1]
CPCS1+2	Copenhagen Prostate Cancer Study 1 and 2	Denmark	Case-control, Clinic-based
EPIC	European Prospective Investigation Into Cancer and Nutrition	EU*	Nested Case-control, prospective cohort[1]
EPIC-Norfolk	European Prospective Investigation of Cancer - Norfolk	UK	Nested Case-control, prospective cohort[1]
ESTHER	Epidemiological investigations of the chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population	Germany	Case-control, Clinic-based[1]
FHCRC	Fred Hutchinson Cancer Research Centre	USA	Case-control, Population-based[1]
MAYO	Mayo Clinic	USA	Case-control, Clinic-based[3,4]
MCCS	Melbourne Collaborative Cohort Study	Australia	Nested Case-control, prospective cohort[2]
MEC	Multiethnic Cohort Study	USA	Nested Case-control, prospective cohort
MOFFITT	The Moffitt Group	USA	Case-control, Clinic-based[1]
PCMUS	Prostate Cancer study Medical University Sofia	Bulgaria	Case-control, Clinic-based[1]
Poland	The Poland Group	Poland	Case-control, Population-based
ProtecT	Prostate testing for cancer and Treatment	UK	Case-control, PSA screening trial[1]
QLD	Retrospective Queensland Study (QLD) and the Prostate Cancer Supportive Care and Patient Outcomes Project (ProsCan)	Australia	Case-control, Clinic-based
SEARCH	Study of Epidemiology and Risk factors in Cancer Heredity	UK	Case-control, Population-based[1]
STHM1	Stockholm 1	Sweden	Cohort[5]
TAMPERE	Finnish Genetic Predisposition to Prostate Cancer Study	Finland	Case-control, Clinic-based+Population-based screening trial
UKGPCS	U.K. Genetic Prostate Cancer Study and The Prostate Cancer Research Foundation Study	UK	Case-control, Clinic-based[1,2,3]
ULM	Institut fr Humangengetik Ulm	Germany	Case-control, Clinic-based[1,3]
UTAH	UTAH Study	USA	Case-control[3]

*Multicenter study from Denmark, Germany, Greece, Italy, Netherlands, Spain, Sweden and UK

1. Controls frequency-matched to cases by five-year age groups.
2. Cases enriched for early disease onset. UKGPCS: age<60, MCCS: age<55.
3. Cases enriched for family history of prostate cancer.
4. Cases enriched for aggressive disease.
5. Men who had undergone a biopsy in the Stockholm area between 2005-2007.

of studies in PRACTICAL are case-control studies, either population-based (cases identified via regional or national population registers) or clinic-based (prostate cancer patients ascertained at oncology clinics). Furthermore, some of the studies are nested case-controls studies within prospective cohorts. Controls were usually selected from the general population in the region of the particular study, frequency matched on age (5 year intervals) to the cases.

4.3 CONOR

The Cohort of Norway (CONOR)[118] was used for replication of top findings in **Study IV**. This is a population-based cohort; ten different surveys have provided data from various parts of Norway, both from urban and rural areas. Approximately 180,000 Norwegians have answered health questionnaires, undergone physical examination and donated non-fasting blood samples (from which DNA was extracted). A record linkage between CONOR and the Norwegian Cancer Register performed in February 2014 identified 4,923 men with a diagnosis of prostate cancer. After exclusion of men diagnosed prior to inclusion in CONOR or after age 80 years 3,614 prostate cancer cases remained. A case-cohort design was employed among the remaining prostate cancer cases to investigate inherited genetic causes of cause-specific survival. From this prostate cancer cohort a random sub-cohort of 931 patients (of which 226 had died due to prostate cancer during follow-up), and all remaining 565 patients that had died from prostate cancer were selected. These individuals were followed for 8741.7 person-years (min: 0.08, median: 5.8, max: 20.8). Genotyping was performed using TaqMan assays. CONOR has recently joined the PRACTICAL consortium and will be a part of the OncoArray genotyping.

4.4 SWEDISH REGISTERS

In **Study II**, we used the unique national registration number that all Swedish citizens have, to perform record linkage between three national registers: the Cancer Register, the Multi-generation Register and the Cause of Death Register. This was done to identify brother pairs and father-son pairs with prostate cancer, which were followed for prostate cancer survival.

4.4.1 Cancer Register

The national Swedish Cancer Register was established in 1958 and is maintained by the National Board of Health and Welfare (Socialstyrelsen)[119]. Since clinicians are obliged by Swedish law to report new cancer diagnosis to the Cancer Register the coverage of the whole population is nearly complete (in a quality study of the register 3.7% of all cancers were underreported[120]) . Tumors are recorded according to the seventh version of the International Classification of Disease (ICD7), where the diagnosis code for prostate cancer is 177. Moreover, the register holds patient information (sex and place of residence) and medical records (date and age at diagnosis, the clinic where the diagnosis was established and histopathological diagnosis, which almost exclusively is adenocarcinoma for prostate cancers). Additionally, the register contains information regarding TNM stage of prostate

cancer tumors since 2002. The data extraction for our analysis of the Cancer Register included data through until the end of year 2009.

4.4.2 Multi-generation Register

The Multi-generation Register contains parental records of all individuals that were born after 1932 and that were registered in any census after 1961[121]. The register contains 7.7 million Swedish born individuals, 97% have a mother and 95% have a father registered. The parental information for offspring born after 1961 is nearly complete. The completeness of individuals born outside of Sweden is considerably lower (27% mothers and 22% fathers are registered)[121]. However, we only included prostate cancer patients born in Sweden in our study. With the use of this register we could identify 1,728 brother pairs and 6,444 father-son pairs where both family members were diagnosed with prostate cancer.

4.4.3 Cause of Death Register

The Cause of Death Register contains ICD coded death causes of all deceased individuals in Sweden. The register is maintained by the National Board of Health and Welfare (Socialstyrelsen). Information from the early years of the register (1952-1960) has less optimal coverage. From 1961 the general quality is excellent and almost complete registration of vital status exists since 1997. The frequency of missing death certificates in the register has increased from 0,006% in 1975 to 0,8% in 2008[122]. When defining prostate cancer-specific cause of death, the underlying cause was used, whereas contributory causes were not considered. Our copy of the register was updated through to 31 December 2010. We followed individuals from their date of prostate cancer diagnosis to death or end of follow-up. Furthermore, we stopped following individuals 20 years after diagnosis to avoid influential outliers and after age 90 years, since the quality of cause of death registration decreases with increasing age. Individuals were considered as events in the survival analysis if they had died from prostate cancer, otherwise they were censored.

4.5 PROCAP

The PROCAP cohort was used in **Study III** to analyze the association between 23 established prostate cancer risk SNPs and disease progression among patients with localized disease. The cohort was recruited from the National Prostate Cancer Register (NPCR), including 98% of all prostate cancer patients reported to the Cancer Register. Information about TNM stage, Gleason score, serum PSA levels at diagnosis and primary treatment within 6 months after diagnosis were available from the NPCR[123,124]. The data in NPCR has excellent quality[125] and is representative of all men with prostate cancer in Sweden[126]. Individuals that were registered with a localized prostate cancer in the NPCR between January 1, 1997 (January 1, 1998 in one region) and December 31, 2002 were included in a retrospective nationwide cohort study, called NPCR of Sweden Follow-Up Study[127,128]. Participants who fulfilled the following criteria were eligible: 70 years or younger at the date of diagnosis, diagnostic serum PSA levels of less than 20 ng/ml; local tumor stage T1-T2;

and no signs of lymph node metastasis (NX or N0) or bone metastasis (MX or M0). In total, 7,960 out of 8,304 eligible (96%) accepted inclusion in the study.

In year 2007, all patients still alive in the NPCR of Sweden Follow-Up study were invited to an extended study, which aimed to assess the importance of genetic and life-style factors on the outcome of localized prostate cancer (PROCAP, PROgression in Cancer of the Prostate). In total 5,431 (77%) of all 7,074 eligible patients accepted inclusion to the PROCAP study by donating a blood sample for DNA extraction and completed a questionnaire regarding life-style factors and physical activity. A total of 529 individuals with unknown primary treatment were excluded from the study since definition of disease progress is dependent on patient's primary treatment. Furthermore, in this study we only included patients who were initially curatively treated (radical prostatectomy or radiation therapy) or were on surveillance (active surveillance or watchful waiting). In total, 3,514 men treated with curative intent and 1,159 patients on surveillance were assessed for disease progression.

Information regarding prostate cancer progression was extracted from medical records at a median time of 4 years after the date of diagnosis. Extracted information included subsequent PSA testing, signs of local progress and distant metastases, reasons for and date of termination of surveillance, and date of last follow-up. Biochemical recurrence was defined according to the primary treatment regimen. A doubling in PSA above the post treatment nadir and exceeding at least 1 ng/ml defined biochemical recurrence among patients who underwent radiation therapy. For patients treated with radical prostatectomy, biochemical recurrence was defined by two consecutive PSA measurements above 0.2 ng/ml. Date of recurrence was set to the date of the first of these two test occasions. Furthermore, operated patients with only one registered PSA test value above 0.5 ng/ml were also considered as having a biochemical recurrence. For patients treated with curative intent, disease progression was defined as a composed event reflecting biochemical recurrence, local progress, or distant metastases. For patients on surveillance, disease progression was defined by the event of termination of deferred treatment with biochemical progression as reason for termination. Date of prostate cancer progression was defined as the earliest date observed for each treatment specific definition of progressive events. Patients without disease progression were censored at last date of follow-up. Patients on surveillance that chose to end deferred treatment without any signs of progress were censored at the date of termination.

The PROCAP participants were genotyped for 23 established prostate cancer susceptibility SNPs, all established SNPs known at that time, using a MassARRAY QGE iPLEX system (Sequenom, Inc. San Diego, CA). The concordance rate among duplicated control samples was 100% and average genotype call rate was 99.7%. Each of the SNPs on the autosomal chromosomes was in Hardy-Weinberg equilibrium ($P \geq 0.01$). We excluded 46 individuals due to low genotyping success rate (<95%).

5 METHODS

5.1 PREDICTION MODELS

5.1.1 Variable selection

In **study I** we aimed to optimize a prediction model for prostate cancer incidence based on the 85,278 SNPs from the iCOGS chip, suggested to be relevant for prostate cancer. This could be viewed as a high-dimensional classification problem, where we want to classify *N* individuals as prostate cancer patients or healthy controls, using a set of *p* predictors (SNPs), where *p>N*. A large number of these predictors are redundant (multicollinear variables), false positives or totally unrelated to the outcome, which introduces "noise" to a prediction model and reduces the predictive capacity. Thus, the exclusion of non-predictive variables is crucial for prediction performance.

There are three main methods for variable selection in prediction modelling: *wrappers, embedded methods* and *filters*[129]. Greedy forward and backward selection procedures, often combined with cross-validation (described below) are examples of wrappers. Penalized regression models (such as Lasso, Elastic Net and Ridge) are examples of embedded methods, where the variable selection is intrinsic in the model[130]. These models contain penalty functions that produce biased estimates of the predictors and shrink some regression coefficients to zero, which in theory should result in better predictions. We evaluated some of the mentioned methods in our data, but in our final prediction model a filtering approach was employed for variable selection. Predictors were ranked based on their univariate association (p-values) with the outcome (case or control status) and included stepwise in the prediction models in the ranked order. Since we wanted to exclude redundant predictors, only the top associated SNP from each Linkage Disequilibrium (LD) block (described in **section 5.4.1**) was included in the model.

5.1.2 Polygenic risk score model

The combined effects of SNPs were incorporated in our prediction models via a polygenic risk score (PRS), which assumes that SNP effects are independent and log-additive, i.e. no interaction between different loci[131]. For each individual *j*, the polygenic risk score for *N* genetic variants was calculated as a weighted sum of the number of risk alleles:

$$\mathrm{PRS} = \sum_{i=1}^{N} w_i * n_{ij}, \qquad (1)$$

where w_i is the effect (logarithm of the per allele odds-ratio) of a SNP at locus *i*, and n_{ij} is the number of risk alleles carried (0, 1 or 2 for autosomal SNPs and 0 or 1 for SNPs on chromosome X). We constructed two risk scores in our prediction model, one that included 65 previously established risk variants (PRS_1) and one where novel SNPs were added (PRS_2).

The final model that we optimized was a logistic regression model with PRS_1 and PRS_2 as covariates:

$$\text{logit}(\text{PC}) = \beta_0 + \beta_1 * \text{PRS}_1 + \beta_2 * \text{PRS}_2, \quad (2)$$

where $\text{logit}(\text{PC})$ is the log-odds of being a prostate cancer case and $\beta_i's$ are regression coefficients. The non-established SNPs were added to PRS_2 according to the filtering strategy described above. In order to optimize the prediction of model (2), the number of SNPs to include in PRS_2 and the weights (w_i) in both risk scores were tuned by a cross-sample validation.

5.1.3 Internal and external validation

If we would optimize our prediction model in the whole PRACTICAL sample, the model would predict the outcome over-optimistically in the same data. However, the prediction performance in an independent sample would be worse, since the model was optimized to fit the data it was trained in[130]. This well-known phenomenon in prediction modeling is called over-fitting. Thus, it is important to validate the model in external test data that was not used to develop (train) the model. We used individuals from SEARCH, one of the populations in PRACTICAL, for external validation. The rest of the PRACTICAL studies were used as training data.

A strategy to overcome the problem of over-fitting in the training data is to use cross-validation[130]. A common approach is to do k-fold cross-validation, where the sample is divided into k random folds that are used for internal validation. Since we want to illustrate that prediction performance varies between studies (with different study designs and genetic composition of the populations), we used different sub-studies in PRACTICAL for internal validation instead, which is a form of cross-study validation. We selected large studies (some studies located geographically close were merged to increase the sample size) for internal validation: Australia (MCCS+QLD), Denmark (CPCS I+II), ProtecT, UKGPCS, USA (FHCRC+MAYO) and STHM1. The first step of the cross-study validation is to set aside one of the internal validation samples. The rest of the data is used to rank SNPs according to the filtering strategy described above and to estimate weights (w_i), which are then used to construct prediction models with varying number of genetic variants in PRS_2. The prediction performances of these models are evaluated in the sample which was set aside. This is executed for each internal validation sample.

5.1.4 Prediction performance

We assessed the prediction performance in the cross-sample validation by measuring how well the model separates the outcome categories (prostate cancer cases vs healthy controls) for various numbers of predictors (SNPs) in PRS_2. For a dichotomous outcome a measure of discrimination can be derived from the receiver operating characteristics (ROC) curve. For example, let us assume that we want to evaluate a prediction model (2) with a set of *G* SNPs,

and estimated risk score weights ($\widehat{w}_i$) and regression coefficients ($\hat{\beta}_i$) from training data. Based on that model, we can calculate a predicted probability of being a prostate cancer case, $\hat{p}(G)$ for each individual in the validation dataset, which was not used to train the model. Furthermore, we classify individuals as cases or controls based on some critical threshold C:

$$\widehat{\mathrm{Y}}_{\mathrm{i}} = \begin{cases} \mathrm{Case} & \text{if } \hat{\mathrm{p}}(\mathrm{G}) \geq \mathrm{C} \\ \mathrm{Control} & \text{if } \hat{\mathrm{p}}(\mathrm{G}) < \mathrm{C} \end{cases}. \tag{3}$$

For a particular threshold C we can then obtain the proportion of correctly and falsely classified cases (sensitivity and 1-specificity). For each possible value of C, we can plot the sensitivity on the y-axis against 1-specificity, which results in a ROC curve. **Figure 6** is an example of a ROC curve, which assess the prediction accuracy of our final model on the external test data (SEARCH).

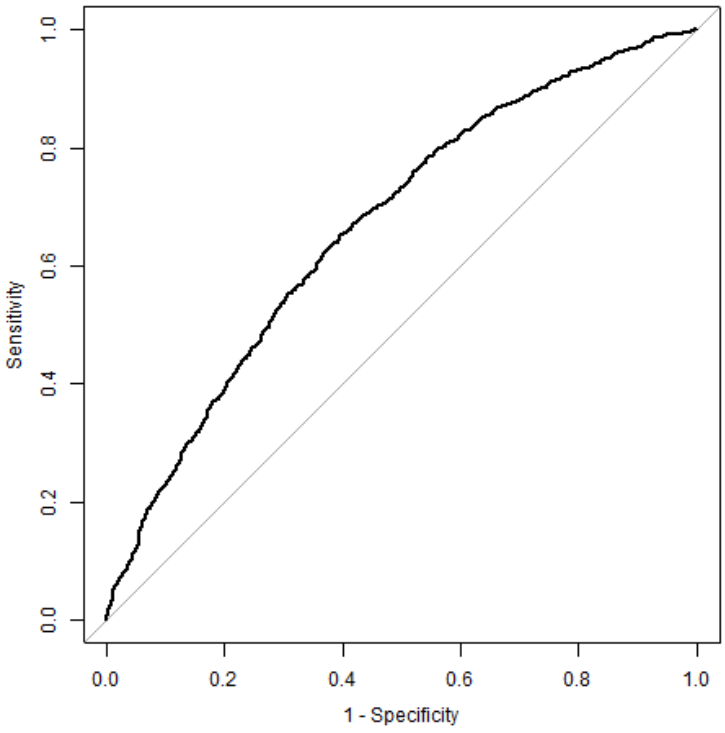

Figure 6: *The black line is a ROC curve of the sensitivity (proportion of correctly classified cases) vs 1-Specificity (proportion falsely classified cases) for different cut-off values (C) in a prediction model. The straight grey line on the diagonal corresponds to a scenario where the model has no predictive capacity (equivalent to coin tossing).*

The area under the ROC curve, AUC (sometimes also called the c-statistic) is a measure of prediction performance of the model. The AUC could be interpreted as the probability that a model classifies a randomly selected pair of a case and a control correctly[132,133]. A model with AUC=1 discriminates cases and controls perfectly and a completely uninformative model (i.e. equivalent to tossing a coin) would have an AUC=0.5. In order to compare the prediction

performance between two models, the difference in AUC could be tested with DeLong's test. This is essentially an extension of a generalized Mann-Whitney test[134,135].

A common critique of the AUC measure is that it is not sensitive to important clinical improvements. In order to achieve a statistical significant change in AUC it has been shown that enormous odds ratios of new predictors are required[136,137]. As a result of this critique, Pencina and coworkers proposed a different measure, called the Net Reclassification Index (NRI)[138]. This measure assesses how cases and controls are reclassified when a prediction model is developed (by adding SNPs in our study). We define an upward (up) reclassification as moving from being classified as a healthy control to a prostate cancer case when SNPs are added to the model and a downward (down) reclassification as moving from case status to control status. NRI consists of a sum of two parts: the net proportion of cases that is correctly reclassified upwards and the net proportion of controls that is correctly reclassified downwards:

$$\mathrm{NRI} = [\mathrm{P}(\mathrm{up}|\mathrm{Case}) - \mathrm{P}(\mathrm{down}|\mathrm{Case})] + [\mathrm{P}(\mathrm{down}|\mathrm{Control}) - \mathrm{P}(\mathrm{up}|\mathrm{Control})]. \quad (4)$$

In **Study I,** NRI was used in the external test data (SEARCH) and statistical tests (based on McNemar's test) were used to assess significant reclassification.

5.2 SURVIVAL ANALYSIS

The primary outcome of interest in **Studies II-IV** was time from prostate cancer diagnosis to progress or death from the disease, with censoring due to end of follow-up. In studies with right censored data, survival analysis models are indispensable and definitions of the survival- and hazard functions are fundamental. If we let *T* be the time to an event, the survival function is defined as the probability to survive longer than a given time *t*:

$$\mathrm{S}(\mathrm{t}) \stackrel{\mathrm{def}}{=} \mathrm{P}(\mathrm{T} > \mathrm{t}). \quad (5)$$

The hazard function is defined as

$$\mathrm{h}(\mathrm{t}) \stackrel{\mathrm{def}}{=} \lim_{\Delta \mathrm{t} \to \infty} \frac{\mathrm{P}(\mathrm{t} \leq \mathrm{T} < \mathrm{t} + \Delta \mathrm{t} | \mathrm{T} \geq \mathrm{t})}{\Delta \mathrm{t}}, \quad (6)$$

which is interpreted as the instantaneous event (for example, prostate cancer death) rate for those who are still at risk at time t. A relationship between these two quantities is given by:

$$\mathrm{S}(\mathrm{t}) = \exp\left[-\int_0^{\mathrm{t}} \mathrm{h}(\mathrm{u})\mathrm{du}\right]. \quad (7)$$

5.2.1 Cox proportional hazards model

The Cox proportional hazards model is one of the most frequently used models in biomedical research[139]. We assume that we have censored data $(T_1,\ \delta_1,\ \boldsymbol{x_1})$, …, $(T_N,\ \delta_N,\ \boldsymbol{x_N})$ with

observed times T_j, a set of observed covariates x_j and a variable δ_j, which indicates if the observation is an event ($\delta_j = 1$) or censored ($\delta_j = 0$). Given a set of covariates $\boldsymbol{X}$ the Cox regression model is defined as

$$\mathrm{h}(\mathrm{t}|\mathbf{X}) \stackrel{\text{def}}{=} \mathrm{h}_0(\mathrm{t}) * \exp(\boldsymbol{\beta}\mathbf{X}), \tag{8}$$

where $h_0(t)$ is a baseline hazard function and $\boldsymbol{\beta}$ are regression coefficients. The model is assumed to be semi-parametric since the baseline hazard function does not need to have a parametric form. Furthermore, the model assumes proportional hazards, i.e. the hazard functions of two groups with different covariate levels ($\boldsymbol{X}$) are proportional. If this is not true then the estimates of the model will be biased.

5.2.2 Partial likelihood function

The Cox partial likelihood is constructed to estimate the $\boldsymbol{\beta}$ coefficients with standard errors[139,140]. Given a sample with ordered event times $T_1 < T_2 < \cdots < T_N$ the partial likelihood of the Cox model is

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{\mathrm{j}=1}^{\mathrm{N}} \left(\frac{\mathrm{e}^{\mathbf{x_j}\boldsymbol{\beta}}}{\sum_{\mathrm{k}\epsilon \mathrm{R_j}} \mathrm{e}^{\mathbf{x_k}\boldsymbol{\beta}}} \right), \tag{9}$$

where j is an index for the event times and R_j is the risk set, which consists of individuals still alive at event time T_j. A censored observation will only contribute to R_j prior to being censored. In this likelihood we can observe that only the order of observation times matter, we do not have to make any assumptions regarding the baseline hazard[141]. In **Study IV**, participants were not always included in the study directly at prostate cancer diagnosis. Thus, they were actually not a part of the risk set until they were recruited. This is an example of left truncated data. The risk sets R_j in Equation (9) is then adjusted for delayed entry, where individuals are at risk if they enter before time T_j and are censored or have an event after T_j.

The maximum likelihood estimations (MLE) $\widehat{\boldsymbol{\beta}}$ are obtained from the solution of the score equations

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{logL}(\boldsymbol{\beta}) = 0. \tag{10}$$

The variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ is estimated from the *observed Fisher information*

$$\mathrm{I}(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \mathrm{logL}(\boldsymbol{\beta}), \tag{11}$$

evaluated at the MLE $\widehat{\boldsymbol{\beta}}$ and the standard errors are calculated from the square root of its diagonal elements.

5.2.3 Cox model in case-cohort studies

The Cox model is commonly used in prospective full cohort studies. However, with proper adjustments, the model can also be implemented in case-cohort studies. This implementation is originally described in terms of the counting process version of the Cox model[142,143]. We introduce the following notation for the *j*th individual: (i) $N_j(t)$ is the number of observed events up to and including time t, (ii) $Y_j(t)$ is a process which indicates if the individual is at risk (=1) or not (=0) at time t, (iii) $Z_j(t)$ is vector of covariates, which in our case is constant with respect to time, and (iv) $r_j(t) = e^{\boldsymbol{\beta}' \boldsymbol{Z_j(t)}}$ is the subject's risk score. For a full cohort with size n_c, the counting process version of the score equations (10) for the Cox model, are written as

$$U(\beta, t) = \sum_{j=1}^{n_C} \int_0^t \{Z_j(s) - \bar{Z}(\beta, s)\} dN_j(s), \tag{12}$$

where $\overline{\boldsymbol{Z}}(\beta, t)$ is the weighted mean

$$\bar{Z}(\beta, t) = \frac{\sum_{i=1}^{n_C} Y_j(t) r_j(\beta, t) Z_j(t)}{\sum_{i=1}^{n_C} Y_j(t) r_j(\beta, t)}. \tag{13}$$

If all individuals are included from the case-cohort sample in equation (12) this would result in a biased estimate for the full cohort. Instead we adjust for the sampling design by calculating the weighted mean based only on individuals in the randomly selected sub-cohort.

5.2.4 Accelerated Failure Time (AFT) models

One reason why the Cox model is widely used is that no assumptions regarding the shape of the hazard are required. Instead of modeling the hazard function as a function of covariates $\boldsymbol{X}$, a parametric model of the survival function could be used instead. In **Study II** we assume that the observed survival times have a log-normal shape. Since we analyze prostate cancer survival within family pairs in that particular study, we also introduce a family index *i*. This gives the following model:

$$T_{ij} = \boldsymbol{\beta}^T \mathbf{X}_{ij} + \varepsilon, \tag{14}$$

where T_{ij} is the log survival time for individual *j* in family *i*, and ε is a normal variable with mean 0. The implication of this model is that the effect of a unit change in covariate *j* acts multiplicatively on the event times with an acceleration factor $e^{\boldsymbol{\beta_j}}$.

5.3 QUANTITATIVE GENETICS

5.3.1 Definition of heritability (narrow-sense)

Quantitative genetics theory could be used to estimate the heritability of a trait (in our case prostate cancer survival). The underlying assumption in quantitative genetics is that each individual's phenotypic outcome (P) is a result of genetic (G) and environmental factors (E)[144], that is:

$$\mathrm{P} = \mathrm{G} + \mathrm{E}. \tag{15}$$

The genetic effect could be further decomposed into an additive component (A) and a dominance component (D) and gene-gene interactions in different loci (epistasis). For simplicity epistasis is usually ignored. This gives the following model:

$$\mathrm{P} = \mathrm{A} + \mathrm{D} + \mathrm{E}. \tag{16}$$

Additive genetic effect assumes that an average effect is added for each copy of a particular allele (i.e. an individual with two copies of the same allele has twice the effect on the phenotype as an individual with one copy) at each locus. Furthermore, the effects in different loci are assumed to act independently and additively. A dominance effect reflects a situation where having one or two copies of an allele results in equal effects.

In general, we are not interested in a particular individual's phenotype, we rather want to model the variability in prostate cancer survival, that is why some die closer to their diagnosis while others do not. Thus, we look at the variance of the phenotype, and by using probability theory we get:

$$\sigma_{\mathrm{P}}^2 = \sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{E}}^2 + 2 * \sigma_{\mathrm{AD}} + 2 * \sigma_{\mathrm{AE}} + 2 * \sigma_{\mathrm{DE}}, \tag{17}$$

where σ_X^2 is the variance of X and σ_{XY} is the covariance of X and Y. This expression is reduced since the additive and dominant genetic effects are independent by definition, i.e. $\sigma_{AD} = 0$. Furthermore, we simplify reality by assuming that no gene-environment interactions exist ($\sigma_{\mathrm{AE}} = 0$ and $\sigma_{\mathrm{DE}} = 0$), which gives the following variance model:

$$\sigma_{\mathrm{P}}^2 = \sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{E}}^2. \tag{18}$$

Moreover, the environmental component is usually decomposed into a part that is shared by individuals within a family (C) and one that is not shared (E), the latter of these also include measurement error. Thus, given the model assumptions mentioned, the total variance of the phenotype can be decomposed as:

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2. \quad (19)$$

Heritability is defined as the proportion of the total phenotypic variance which is attributable to genetics[145]. However in most applications, *narrow-sense heritability, h^2* is estimated, which is defined as the proportion of the total phenotypic variation that is explained by additive effects:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}. \quad (20)$$

We estimated this quantity, which will be referred as *heritability*, for prostate cancer survival in **Study II**.

5.3.2 ACE model

Family relations, i.e. covariance structures between family members can be used to estimate the heritability h^2. Traditionally, this is done in twin studies[146,147], where monozygotic (MZ) twins are compared with dizygotic (DZ) twins, because they share different amount of genes but the same environment. However, other family structures are also used to estimate the relative balance between genetic and environmental effects. It is possible to use families where the amount of genetic sharing is equal but the environment within a family is different. An example is adoption studies where siblings that are reared together are contrasted to siblings that are reared apart. In our study we estimate the heritability of prostate cancer survival in families that are genetically equal (share 50% of segregating alleles identical by descent), brother pairs and father-son pairs. However, since the environment that affect prostate cancer prognosis is expected to change with calendar time, we assume that brothers share environment (since they get their diagnosis closely in time), whereas fathers and sons do not. The latter is probably not entirely accurate but with the family structures that we have in our data this assumption is necessary to be able to estimate heritability.

With our data structure we can only study three components of the right hand side of equation (19) at the same time. The unshared environmental component (E) has to remain in the model since random errors are included in that parameter. Furthermore, it is common to model the additive genetic (A) and shared environmental (C) components, unless it is believed that dominance effects are important for the trait (which we have no reason to believe in case of prostate cancer survival). This leaves us with a so called ACE-model.

We model the outcome in our study, log-time from prostate cancer diagnosis to death from prostate cancer, T_{ij} (*i*th family for the *j*th family member) in a mixed log-normal AFT model:

$$T_{ij} = \mu_{ij} + A_{ij} + C_i + E_{ij}, \quad (21)$$

with unmeasured latent variables A_{ij}, C_i and E_{ij} (additive genetic, common environment and unshared environment components) that are assumed to be normally distributed with mean 0 and variances $\boldsymbol{\varphi} = (\sigma_a^2, \sigma_c^2, \sigma_e^2)$ respectively. The fixed effects, $\mu_{ij} = \boldsymbol{\beta X_{ij}}$ are the observed variables. We adjusted our analysis for age and calendar period of diagnosis. The calendar period effect on prostate cancer survival was modeled as linear up to 1995 with a quadratic term after that (because of the introduction of PSA testing; see **section 2.1.7**).

Given the previously mentioned assumptions, the expected covariance matrices of the prostate cancer log survival times are

$$\mathrm{Cov}\begin{pmatrix}\mathrm{T_{i1}}\\ \mathrm{T_{i2}}\end{pmatrix} = \begin{bmatrix}\sigma_a^2 + \sigma_c^2 + \sigma_e^2 & 0.5 * \sigma_a^2\\ 0.5 * \sigma_a^2 & \sigma_a^2 + \sigma_c^2 + \sigma_e^2\end{bmatrix}, \tag{22}$$

for father-son pairs, and

$$\mathrm{Cov}\begin{pmatrix}\mathrm{T_{i1}}\\ \mathrm{T_{i2}}\end{pmatrix} = \begin{bmatrix}\sigma_a^2 + \sigma_c^2 + \sigma_e^2 & 0.5 * \sigma_a^2 + \sigma_c^2\\ 0.5 * \sigma_a^2 + \sigma_c^2 & \sigma_a^2 + \sigma_c^2 + \sigma_e^2\end{bmatrix}, \tag{23}$$

for brothers.

We let the observed log-transform of the survival times, $\boldsymbol{T_i} = [T_{i1}\ T_{i2}]^T$ in family pair *i* have a bivariate normal distribution with mean $\boldsymbol{\mu} = [\mu_1\ \mu_2]^T$ and covariance

$$\Sigma = \begin{bmatrix}\sigma_1^2 & \rho_i\sigma_1\sigma_2\\ \rho_i\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}, \tag{24}$$

where σ_j^2 is the variance of family member *j* and $\rho_i\sigma_1\sigma_2$ is the family pair covariance. We will refer to this as the observed covariance matrix. Furthermore, we assume that $\boldsymbol{T_i}$ are independent from the log censoring times $\boldsymbol{C}_i = [C_{i1}\quad C_{i2}]^T$. This leads to the following definition of the observed prostate-specific survival times, $\boldsymbol{Y}_i$:

$$\mathbf{Y_i} = \begin{bmatrix}\mathrm{Y_{i1}}\\ \mathrm{Y_{i2}}\end{bmatrix} = \begin{bmatrix}\min(\mathrm{T_{i1}}, \mathrm{C_{i1}})\\ \min(\mathrm{T_{i2}}, \mathrm{C_{i2}})\end{bmatrix}. \tag{25}$$

Thus, an event is not observed if $\mathrm{C_{ij}} < \mathrm{T_{ij}}$ and the exact survival time is not known (but it is known that $\mathrm{T_{ij}} > \mathrm{C_{ij}}$).

Moreover, we assume that variances (adjusted for age and calendar time) within family pairs are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). This implicates that the observed covariance matrix (24) simplifies to σ^2on the main diagonal and to the covariance elements $\sigma^2\rho_{FS}$ and $\sigma^2\rho_B$ in fathers-son pairs and brother pairs, respectively. Hence, if we compare these observed covariance matrices with the expected, (22) and (23), we can see that the variance components $\boldsymbol{\varphi}$ are identified from the following equations:

$$\sigma_a^2 + \sigma_c^2 + \sigma_e^2 = \sigma^2,$$
$$0.5 \cdot \sigma_a^2 + \sigma_c^2 = \sigma^2 \rho_B,$$
$$0.5 \cdot \sigma_a^2 = \sigma^2 \rho_{FS}. \quad (26)$$

Since we have three unknown parameters and equally many equations only one unique solution exist. Thus, it is clear that including a fourth variance parameter in the model, for example a dominance genetic component (D) would result in an unidentifiable problem.

5.3.3 Conditional likelihood estimation

The following two sections are adapted from the appendix of **Study II**. In order to obtain estimates of $\boldsymbol{\varphi}$ we want to maximize the sum of the log-likelihood contributions of all pairs with regards to the unknown parameters, $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_i)$:

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \log(L_i(\boldsymbol{\theta})). \quad (27)$$

The likelihood components for the family pairs are calculated depending on whether the events are observed or not. There are four possible scenarios, shown in **Figure 7**, for the censoring in a family pair, which we have to condition on: (i) both survival times are observed as prostate cancer events (no censoring); (ii) only member 1 is censored; (iii) only member 2 is censored; and (iv) both members are censored. For censored observations we want to integrate the density function over all possible survival times, i.e. from $\boldsymbol{C}_i$.

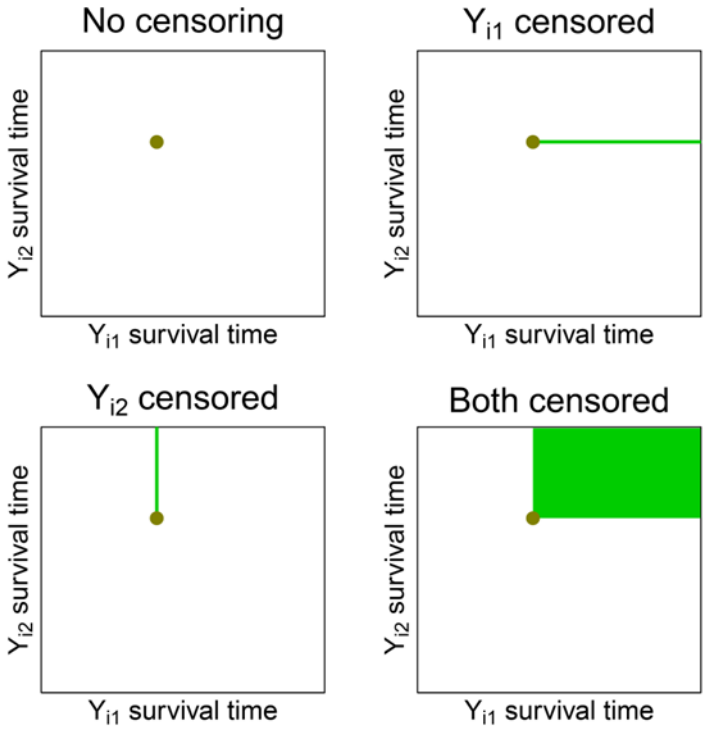

Figure 7: *Four scenarios of censoring of survival times in a family pair. The dot represents observed times, censored or not, the green lines and area represent potential survival times if censored.*

If both individuals are observed events in the *i*th family with survival times $\boldsymbol{y_i} = (t_{i1}\ t_{i2})^T$, the likelihood is given by the multivariate normal density function:

$$L_i(\theta) = f_{Y_1,Y_2}(\mathbf{y_i}) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y_i}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y_i}-\boldsymbol{\mu})\right)$$

$$= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}\exp\left(-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\left(\sigma_2^2(y_{i1}-\mu_1)^2+\sigma_1^2(y_{i2}-\mu_2)^2 - 2\rho\sigma_1\sigma_2(y_{i1}-\mu_1)(y_{i2}-\mu_2)\right)\right). \quad (28)$$

In the scenario where one of the observations in a pair is censored, for example Y_{i1} at C_{i1}, the observed times are $\boldsymbol{y}_i = (C_{i1}, t_{i2})^T$. The likelihood in this case is constructed by using the conditional density function, where we condition on that we have observed an event (prostate cancer death) for the second family member (i=2):

$$f_{Y_{i1}|Y_{i2}=y_{i2}}(y_{i1}) = \frac{f_{Y_{i1},Y_{i2}}(y_{i1},y_{i2})}{f_{Y_{i2}}(y_{i2})}$$

$$= \frac{\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}\exp\left(-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\left(\sigma_2^2(y_{i1}-\mu_1)^2+\sigma_1^2(y_{i2}-\mu_2)^2-2\rho\sigma_1\sigma_2(y_{i1}-\mu_1)(y_{i2}-\mu_2)\right)\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{1}{2\sigma_2^2}(y_2-\mu_2)^2\right)},$$

$$= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}}\exp\left(\frac{1}{2\sigma_1^2(1-\rho^2)}\left(y_1-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2)\right)^2\right) \quad (29)$$

The last expression is recognizable as a $N\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2), \sigma_1^2(1-\rho^2)\right)$-distributed variable. Thus, the likelihood may be calculated as

$$L_i(\boldsymbol{\theta}) = f_{Y_2}(y_2)\int_{c_{i1}}^{\infty} f_{Y_1|Y_2=y_2}(y_1)dy_1, \quad (30)$$

where $f_{Y_2}(y_2)$ is the density function of a $N(\mu_2, \sigma_2^2)$ distributed variable. This is of course symmetric for the scenario where Y_{i2} is censored and Y_{i1} is not.

In the case where both observations are censored, at C_{i1} and C_{i2}, the likelihood is the multivariate integral

$$L_i(\boldsymbol{\theta}) = \int_{C_{i1}}^{\infty}\int_{C_{i2}}^{\infty} f_{Y_{i1},Y_{i2}}(y_{i1},y_{i2})dy_1dy_2. \quad (31)$$

Maximization of this this composite likelihood was performed in the open source package OpenMX in the program R. The function mxFIMLObjective (with a thresholds option), was used to integrate the likelihood functions above, from the censoring times, $\boldsymbol{C}_{ij}$ (interpreted as "thresholds" in a liability model for twin data with a binary outcome[146]).

5.3.4 Simulations

The above described conditional likelihood estimation procedure was evaluated on simulated data, which was created to closely mimic the observed data. In order to get the same distribution of age and calendar period (year) at diagnosis in the different family pairs, we took a random sample (25,000 families) with replacement from the real data. This also gave a true relation between the number of father-son pairs (N= 19,665) and brother-pairs (N=5,335). Since a realistic scenario is that the survival time decrease with age and increase with calendar period (because of improved treatment and clinical practice), we simulated in each family *i* for individual *j* a log-normal time from diagnosis to death from prostate cancer, T_{ij}:

$$\begin{bmatrix} \mathrm{T_{i1}} \\ \mathrm{T_{i2}} \end{bmatrix} = \begin{bmatrix} \log(\mathrm{T_1}) \\ \log(\mathrm{T_2}) \end{bmatrix} + \mathbf{X},$$

where $\log(T_j) = mean(\log(T_{\text{real data}})) - 0.02 * AGE + 0.1 * YEAR$ and $\boldsymbol{X}$ has a bivariate random normal distribution with mean 0 and covariance Σ, which is generated by the rmvnorm function (mvtnorm-package) in R. The structure of the covariance matrix depends on family type, given by (22) for father-son pairs and (23) for brother pairs.

Furthermore, for each individual, we simulated a time from diagnosis to death from competing risks (T_{ij}^{CR}) by using national life tables from Statistics Sweden for Swedish males in 2010[148]. Mortality rates for each age-group were used to randomly generate an age of death from competing causes, from an exponential function in which the rate was piece-wise constant by age. The msm-package in R was used to implement this.

The time of event was defined as the minimum of the generated times of cause-specific death and competing risks, i.e. $\min(T_{ij}, T_{ij}^{CR})$. If simulated death from prostate cancer occurred before the competing risk this was considered as an event, otherwise the observation was censored. As in the real data, we also stopped following individuals at age 90, after 20 years of follow-up and at end of follow-up (December 31, 2010). Simulation was performed for various combinations of variance components ($\sigma_a^2,\ \sigma_c^2,\ \sigma_e^2$).

5.4 GENOME-WIDE ASSOCIATION STUDIES (GWAS)

In study **IV**, we performed a GWAS of prostate cancer-specific survival. The basic idea in a GWAS is to scan the whole genome for common SNPs (MAF>0.01) that are associated with a trait of interest. This search is agnostic, in the sense that no prior hypothesis is formulated regarding any specific loci or genes to be associated with the outcome. Using this approach low-risk genetic variants are usually identified.

5.4.1 Linkage Disequilibrium (LD)

During meiosis, i.e. when germ cells are produced, DNA from the paternal and maternal chromosomes recombines (at least once per chromosome). Thus, gametes consist of a mixture of DNA chunks from both parents, and alleles that are located physically close are often inherited together and correlated[70]. Measures of correlation (nonrandom association) in a population, *linkage disequilibrium (LD)*, between two SNP alleles (A and B) on different loci depend on the following coefficient:

$$\mathrm{D} = \mathrm{f_{AB}} - \mathrm{f_A f_B}, \tag{32}$$

where f_{AB} is the frequency of individuals that carry both the A and B alleles and $f_A f_B$ is the product of the A and B allele frequencies[149]. If $D = 0$, the alleles are in linkage equilibrium, that is, they are statistically independent. Since *D* is constrained by the allele frequencies other standardized measures are preferred, such as r^2, which is a measure of correlation between two loci:

$$\mathrm{r^2} = \frac{\mathrm{D^2}}{\mathrm{p_A(1-p_A)p_B(1-p_B)}}. \tag{33}$$

Values of this coefficient range from 0 (no correlation) to 1 (perfect correlation). The human genome is made up of haplotype blocks, containing variants in LD with each other. Thus, in a GWAS it is sufficient to genotype one SNP in each such block when testing for association with disease outcome, since this tagSNP will be in high LD with all variants in that region.

5.4.2 Quality Control (QC)

A first important step in a GWAS is quality control (QC) and data cleaning. Individuals were excluded from the study due to signs of poor DNA quality for one of the following reasons:

1. Genotype call rate <95%.
2. Cryptic and false duplicates.
3. Not male (XX or XXY sex genotypes).
4. High or low heterozygosity (±4.89 SD from mean).
5. Related to other study participant.
6. Ancestry outlier.

Furthermore, SNPs with bad genotyping were excluded based on the following criteria:

1. Call rate <95%.
2. Departure from Hardy-Weinberg equilibrium.

An overview of which of these QC criteria that were applied in each study is given in **Table 3**. Explanations of some of these concepts follows.

5.4.2.1 Hardy Weinberg equilibrium

Let us assume that a SNP has alleles *A* and *a* with allele frequencies *p* and *q* respectively. The Hardy Weinberg principle states that the frequencies of the three possible genotypic variants in a population are $f(AA) = p^2$, $f(Aa) = 2pq$ and $f(aa) = q^2$. This theory assumes random mating, no inbreeding, no selective survival factors related to genotype, a large population and no recent mutation[70]. Large deviations from Hardy Weinberg equilibrium (HWE) may indicate poor genotyping of a SNP.

5.4.2.2 Identical by state (IBS)

Two alleles are identical by state (IBS) if they have identical DNA composition and function but do not necessarily come from the same ancestor (i.e. not identical by descent; IBD). For a pair of individuals, the proportion of SNPs that are IBS indicate to what degree they are genetically related. Based on approximately 37,000 uncorrelated SNPs, a matrix with IBS proportions, for all pairs of individuals, was calculated using the GenABEL package[150] in R. Using this information, duplicates and relatives could be identified. For obvious first-degree relatives and cryptic relatives (IBS>0.86), controls were removed from case-control pairs, otherwise the individual with the lowest genotype call rate was removed.

Table 3: *Summary of quality control.*

			Sample QC				**Genotyping QC**			
			Inclusion/exclusion criteria					**Inclusion criteria**		
Study/ Consortium	**No. of cases in file**	**No. of controls in file**	**Minimum sample call rate for inclusion**	**Exclus-ions***	**No. of cases after exclus-ions**	**No. of controls after exclus-ions**	**Genotyping platform**	**SNP Call rate**	**P, HWE**	**SNPs that met QC criteria**
PRACTICAL iCOGS	24198	23994	≥95%	1−8	20219	20440	Custom Illumina Infinium (iCOGS)	≥95%	$\geq 1x10^{-7}$	201598
UK GWAS stage1	1906	1934	≥97%	1, 2, 4-7	1854	1894	Illumina Infinium HumanHap 550 Array	≥95%	$\geq 1x10^{-5}$	541129
UK GWAS stage 2	3888	3956	≥97%	1, 2, 4-7	3706	3884	Illumina iSELECT	≥95%	$\geq 1x10^{-5}$	43671
CAPS 1 GWAS	498	502	≥95%	1, 2, 4-7	474	482	Affymetrix GeneChip 500K	≥95%	$\geq 1x10^{-6}$	369025
CAPS 2 GWAS	1483	519	≥95%	1, 2, 4-7	1458	512	Affymetrix GeneChip 5.0K	≥95%	$\geq 1x10^{-6}$	369610
BPC3 GWAS	2137	3101	≥95%	2,5-7	2068	3011	Illumina Human610 Illumina 610K	≥95%	$\geq 1x10^{-5}$	525766

*Exclusion criteria = 1) XX or XXY; 2) heterozygosity lower/higher than 4.89 SD from mean; 3) low concordance with previous genotypes; 4) false duplicates; 5) cryptic duplicates; 6) relatives; 7) ancestry outliers; 8) UK, CAPS or BPC3 GWAS overlap.

5.4.3 Principal components

Systematic differences in population allele frequencies across geographic regions results in population structures that can be assessed by principal components of SNP data. For example, on European level, a south-north gradient is present along the axis of the first principal component and a west-east gradient along the second principal component[151]. Intra-country differences have also been observed, but are generally modest[152,153].

Population stratification may confound results in genetic association studies. This could be corrected for, by adjusting for principal components in the model. However, it is also important to not over-adjust (i.e. adjust for principal components that are not needed) since the power to detect true associations decrease. We adjusted for six principal components in the association analysis in **Study I**, where PRACTICAL samples from a mixture of samples across the world were used. In **Study IV** we used principal components in a sensitivity analysis of our top findings. Furthermore, principal components were used to exclude genetic outliers.

5.4.4 Imputation

Genetic data is very suitable for imputation of unmeasured genotypes because of the LD block structure of the human genome. Since alleles that are located physically close are often inherited together, missing genotypes could be inferred from neighboring loci. To illustrate the principle of SNP imputation we can consider a simplistic example with one missing SNP in a haplotype, AC?GA. If we observe in a comparison with a reference population that the haplotype ACCGA is the most common, we would make the guess that the missing genotype should be C, since we know that stretches of DNA are inherited together. However, individuals in the reference population with less frequent haplotypes, for example ACTGA might also exist. In this case we would assign the two possible variants probabilities that correspond to the observed frequencies in the reference population. In practice, imputation is carried out in two main steps: (i) phasing the sample genotypes and (ii) alignment of phased data to phased reference haplotypes, coupled with imputation.

The rationale for using imputed data is that such data increase the marker density and the power to detect a true causal association. However, the main benefit is that it enables meta-analysis across different genotyping platforms[154,155]. In **Study IV** a reference panel from the Phase I release of the 1000 genome project (March 2012)[78], consisting of 1,094 individuals from 14 different populations with 17 million SNPs/indels, was used. Imputation was performed in PRACTICAL (each sub study was imputed separately), CAPS GWAS and UKGPCS GWAS, using the software IMPUTE2[156], while MACH[157] was used for imputation of the BPC3 study. This was carried out at Cambridge University by PRACTICAL consortium collaborators as described in Amin Al Olama et al[86]. Furthermore, we obtain estimates of the imputation accuracy of each SNP. These metrics (r^2 in MACH and INFO in IMPUTE2) lie in the range (0,1), where 0 means complete uncertainty and 1 indicates perfect

imputation[154]. Only SNPs with imputation accuracy above 0.75 were included in the final meta-analysis.

5.4.5 Meta-analysis

For each individual study we carried out genome-wide assessment and combined the estimated effects in a meta-analysis .We performed Cox regression analysis with time from prostate cancer diagnosis to death from prostate cancer (or censoring) as outcome. Delayed entry (left truncation) was allowed in the model since some study participants were not enrolled at the time of diagnosis. Dosage of a reference allele was used as exposure, which corresponds to an additive genetic model. The following example of a *C/T* polymorphism illustrates how SNP dosage was calculated. For imputed SNPs, we obtained probabilities p_{CC}, p_{CT} and p_{TT} for the possible genotypes *CC*, *CT* and *TT* for each individual. These were transformed to dosages of a reference allele (for example *C*):

$$\mathrm{dosage_C} = 2 * \mathrm{p_{CC}} + 1 * \mathrm{p_{CT}} + 0 * \mathrm{p_{TT}}. \tag{34}$$

Results from each study were combined in an inverse variance weighted meta-analysis. Assuming that the additive genetic effect β_i with a standard error SE_i was estimated from study *i* for a particular SNP. With weights defined as $w_i = 1/SE_i^2$, the weighted combined effects and standard errors were obtained by:

$$\beta_{\mathrm{Meta}} = \sum_{\mathrm{i}} \beta_{\mathrm{i}}\, \mathrm{w_i} \Big/ \sum_{\mathrm{i}} \mathrm{w_i} \tag{35}$$

$$\mathrm{SE_{Meta}} = \sqrt{1 \Big/ \sum_{\mathrm{i}} \mathrm{w_i}}. \tag{36}$$

These were used to calculate an overall Z statistic, $\mathrm{Z_{Meta}} = \beta_{\mathrm{Meta}}/\mathrm{SE_{Meta}}$, and used to perform an overall test with $\mathrm{P} = 2\Phi(-|\mathrm{Z_{Meta}}|)$, where Φ is the cumulative normal distribution function. This meta-analysis was carried out for all available SNPs over the genome. Furthermore, for each SNP, Cochran's Q-test for heterogeneity between samples was performed. We only considered SNPs where no significant (p>0.05) heterogeneity between study effects was present. The meta-analysis was implemented in the METAL software[158].

5.4.6 Evaluation of genome-wide association results

After performing GWAS it is critical to evaluate the genome-wide distribution of the used test statistic in comparison with the expected distribution by chance (when no genetic associations are present). Deviance from the expected null distribution may be due to an excess of truly associated genetic variants that are in LD[159]. However, it might also indicate poor quality control of the data, inappropriate adjustment for population structure (principal components) or some other technical bias. A Quantile-Quantile (Q-Q) plot, of the observed p-

values against the expected p-values obtained by chance is a visual assessment of deviance. The resulting plot should ideally appear as a straight line on the diagonal, with true associations deviating in the extreme tail of the distribution[160].

Another commonly used metric is the genomic inflation factor λ, which is defined as the ratio between the median of the observed p-values against the median of p-values expected by chance[161]. This is a measure of the excess of false positives. However, it has been shown that the inflation factor scales with sample size[161] and that the inflation factor λ_{1000} for an equivalent study of 1000 cases and 1000 controls is more informative[162]:

$$\lambda_{1000} = 1 + (\lambda - 1) \times \left(\frac{1}{n_{cases}} + \frac{1}{n_{controls}}\right) \Big/ \left(\frac{1}{500}\right), \quad (37)$$

where, in our study, n_{cases} is the number of prostate cancer patients that died from their disease and $n_{controls}$ is the number that did not die (i.e. were censored). Results are usually displayed in a Manhattan plot with p-values ($-\log_{10}$) plotted on the y-axis against their physical position in the chromosome on the x-axis. Since it is assumed that approximately 1 million independent statistical tests are performed in a GWAS it is crucial to adjust for multiple tests (to avoid false positive findings). A Bonferroni correction gives a critical p-value, $p < 5 \times 10^{-8}$, where a SNP is considered as genome-wide significant, i.e. truly associated with the outcome.

In our study, a SNP was considered as an interesting finding which qualified for replication if the following criteria were satisfied:

1. Genome-wide significant in combined Meta-analysis between PRACTICAL and BPC3.
2. The direction of effect in PRACTICAL and BPC3 were equal.
3. Cochran's Q-test for heterogenity was not significant (p>0.05) within PRACTICAL and BPC3.
4. Cochran's Q-test for heterogenity was not significant (p>0.05) between PRACTICAL and BPC3.

Furthermore, a number of sensitivity analyses were performed for all interesting SNPs:

1. Adjustment for six principal components in a full cohort stratified Cox regression analysis to adjust for possible confounding due to population stratification.
2. Adjustment for other possible confounders, such as age, PSA and Gleason score at diagnosis.

The leading SNP (strongest associated with the outcome) in each interesting region was genotyped in a sub-population (UKGPCS1 GWAS) to assess the imputation quality. If the leading SNP was not possible to genotype because of manufacturing reasons, we chose a different genome-wide significant marker in the same region or a surrogate SNP (in high LD with the leading SNP). SNPs that demonstrated a satisfying concordance rate between

imputed and genotyped SNPs (percentage of individuals correctly classified by imputation) were put forward for replication in an independent sample (CONOR).

6 RESULTS AND DISCUSSION

6.1 STUDY I: PREDICTION OF INDIVIDUAL GENETIC RISK TO PROSTATE CANCER

6.1.1 Results

In **Figure 8** we can see that the prediction performance for different internal validation samples in the cross-sample validation varies substantially. AUC values range from 0.64 to 0.69 in the models that only include 65 established risk variants. In most studies, except STHM1, we can observe an increasing trend in AUC when additional, previously not established SNPs are added to the model. However, this initial growth is rather modest (approximately 0.01 change in AUC) and the peak of mean prediction performance is observed at 68 added SNPs.

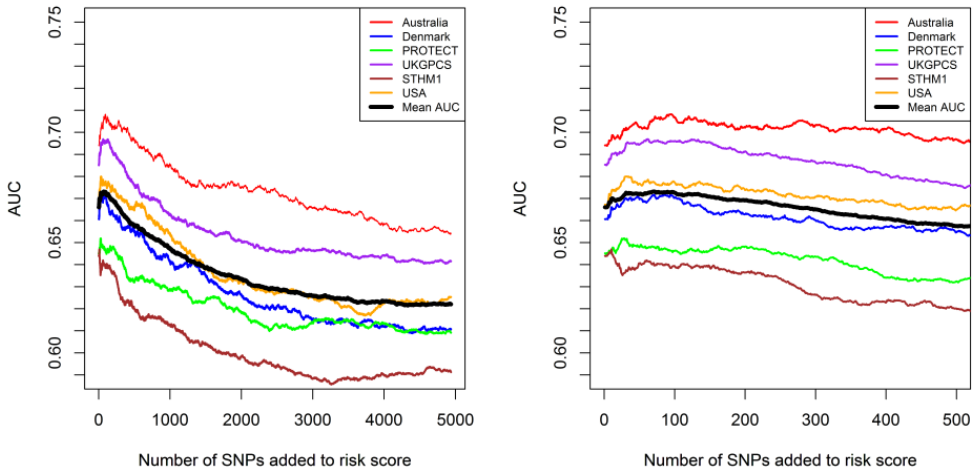

Figure 8: *Prediction accuracy (AUC) for different internal validation samples of the training data. The left plot shows prediction performance when up to 5,000 novel SNPs are added to the prediction model. The right plot is zoomed in on the part where the predictions increase. The black line corresponds to the mean AUC.*

Based on the training data, we concluded that a prediction model with 65 established and 68 additional non-established SNPs would give the best predictive capacity. This final model (with polygenic risk score weights and SNP rank order obtained from the whole training data) was evaluated in the external test sample, SEARCH. The initial model with only 65 established SNPs had an AUC=0.67 and increased significantly (p=0.0012) to 0.68 when the 68 non-established SNPs were added to the prediction model. Furthermore, the NRI was also significantly improved (p=0.00051) when the non-established SNPs were added to the prediction model.

We could also observe an increasing trend in prostate cancer risk with higher total genetic risk score in the independent SEARCH sample (**Table 4**). Comparing with the reference category (40–60% percentiles), individuals with lowest risk scores (lowest 5% percentile)

have an 84% decreased risk of prostate cancer, while individuals with highest risk scores (highest 5% percentile) have a four-fold increased relative risk of prostate cancer.

6.1.2 Discussion

In this study, we derived a polygenic risk score model that discriminates between prostate cancer cases and healthy controls to some extent. We observed a large variability in prediction performance (AUC) in the training data, probably due to different study designs. For example, the best predictions are obtained in the Australian sample where cases with age at diagnosis below 55 were enrolled, followed by the UKGPCS study, which also included early onset cases (below 60 years) and patients with a family history of prostate cancer. These cancer subtypes are known to be more heritable, which results in better predictions by the SNPs in our model. However, we argue that the prediction accuracy (AUC=0.68) in the independent SEARCH study (population based) is a realistic result in a screening situation. The participants in that particular study are aged up to 70 years, and therefore representative for a screening target population. Furthermore, we observed in SEARCH that a polygenic risk score has the capacity to stratify individuals into risk groups based on the 133 SNPs (65 established and 68 novel) in our final model. It is evident from this study that the 5% with highest risk scores has a considerably higher risk of getting prostate cancer compared to the 5% with lowest risk scores. These results are in concordance with other studies with a somewhat different approach[86,163]. However, the clinical utility of these models still remains to be evaluated.

Table 4: *Risk distribution in different percentiles of a genetic risk score, containing both 65 established and 68 novel SNPs, evaluated in an external test sample.*

Percentiles	OR (95%CI)	P value
0% - 5%	0.16 (0.10,0.27)	4.43e-12
5% - 10%	0.52 (0.35,0.77)	0.0012
10% - 20%	0.41 (0.30,0.56)	2.85e-08
20% - 30%	0.82 (0.61,1.10)	0.18
30% - 40%	0.92 (0.69,1.24)	0.60
40% - 60%	1.00 (ref)	-
60% - 70%	1.36 (1.01,1.84)	0.046
70% - 80%	1.60 (1.18,2.16)	0.0026
80% - 90%	2.58 (1.86,3.56)	9.66e-09
90% - 95%	2.37 (1.56,3.60)	5.07e-05
95% - 100%	4.00 (2.51,6.39)	6.50e-09

Furthermore, by adding the SNPs that were mostly associated with the outcome in the training sample to the 65 previously established risk variants, the prediction model was significantly improved. Interestingly, the additional SNPs were all located in regions that were previously recognized as established prostate cancer risk regions. This suggests that fine-mapping of these parts of the genome could identify SNPs that are more predictive (i.e. causal variants or SNPs in higher LD with the causal variants) for prostate cancer incidence. Concurrently with our study, Amin Al Olama and coworkers reported that fine-mapping of 64 known prostate cancer GWAS regions resulted in multiple independent signals in 16 of the regions (12 independent but previously unknown regions within 500kB of the established SNP) and 35 regions with a new lead SNP[94]. This supports our observation that more predictive SNPs exist in previously known GWAS regions.

We employed several other strategies to improve predictions in our polygenic risk score model. We used forward selection, together with cross-validation to rank SNPs in which order to add them into the prediction model. This did not improve the predictive capacity compared with a model that only included established risk SNPs. Furthermore, several regularized regression methods (Lasso, Ridge and Elastic net) and a random forest model were implemented in our data. In general these efforts resulted in over-fitted prediction models that performed poorly in the external test sample (SEARCH).

6.2 STUDY II: ESTIMATING HERITABILITY OF PROSTATE CANCER-SPECIFIC SURVIVAL USING POPULATION-BASED REGISTERS

6.2.1 Results

Results from assessment of the heritability of prostate cancer-specific survival are shown in **Table 5**. No evidence of heritability of prostate cancer-specific survival was observed in the crude unadjusted analysis. However, a model with adjustment for age at diagnosis and calendar period revealed an estimated heritability of 0.10 (95% confidence interval [CI] = 0.00 to 0.20) that was borderline significantly different from zero (P = 0.057). In the adjusted analysis the shared environment was not significantly different from zero with a point estimate of 0.00 (95% CI = 0.00 to 0.13).

Sensitivity analysis, exploring 5- and 10-years survival and excluding PSA detected tumors, revealed that the heritability estimates were relatively robust, varying between 0.10 and 0.14. The estimates of the shared environmental component increased to 0.14 (95% CI = 0.00 to 0.30) when T1c tumors from 2002 were removed, but continued to not be significantly different from zero.

We can see that the observed results follow the same pattern as in the simulated data (**Table 6**). In a crude analysis, heritability estimates are biased towards null and the shared environmental component is clearly over-estimated. However, by adjusting for age and calendar period we obtain precise estimates for the heritability but the shared environmental

component is under-estimated. This underlines the importance of proper adjustment for age and calendar period for the heritability estimates.

Table 5: *Estimates and 95% likelihood-based confidence intervals of heritability-, shared environment- and unshared environment components.*

Model	a2*	c2**	e2***
Main analysis			
Crude (unadjusted)	0.00 (0.00, 0.01)	0.51 (0.42, 0.59)	0.49 (0.41, 0.58)
Age adjusted	0.00 (0.00, 0.03)	0.30 (0.17, 0.42)	0.70 (0.58, 0.83)
Calendar time adjusted	0.05 (0.00, 0.15)	0.02 (0.00, 0.16)	0.93 (0.79, 1.00)
Age and calendar time adjusted	0.10 (0.00, 0.20)	0.00 (0.00, 0.13)	0.90 (0.76, 1.00)
Sensitivity analysis			
10 year survival[a]	0.10 (0.00, 0.20)	0.00 (0.00,0.13)	0.90 (0.77, 1.00)
5 year survival[b]	0.14 (0.02, 0.25)	0.00 (0.00, 0.15)	0.86 (0.71, 0.98)
T1c tumors removed[c]	0.12 (0.01, 0.24)	0.14 (0.00, 0.30)	0.74 (0.58, 0.90)
Diagnosed from 2005 removed[d]	0.13 (0.00, 0.25)	0.01 (0.00, 0.20)	0.87 (0.68, 0.99)

* Estimated heritability, $a2 = \sigma_a^2/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$.

** Estimated common environment effect, $c2 = \sigma_c^2/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$.

*** Estimated unshared environment effect, $e2 = \sigma_e^2/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$.

a, Same data as in main model but observations censored after 10 years.

b, Same data as in main model but observations censored after 5 years.

c, Patients with missing T-stage information and T1c tumors (PSA detected) removed from 2002. Data restricted to 3915 father-son pairs and 746 brother-pairs.

d, Individuals diagnosed from 2005 removed. Data restricted to 2776 father-son pairs and 461 brother pairs.

Table 6: *Estimates and 95% confidence intervals of variance components* ($\boldsymbol{a2}=\boldsymbol{\sigma_a^2}$, $\boldsymbol{c2}=\boldsymbol{\sigma_c^2}$, $\boldsymbol{e2}=\boldsymbol{\sigma_e^2}$) *on simulated data.*

	Crude model (no adjustment)			Adjustment for age and calendar period (year)		
Parameters in simulated data	**a2**	**c2**	**e2**	**a2**	**c2**	**e2**
a2=0.1, c2=0.2, e2=0.7	0.00 (0.00,0.00)	0.76 (0.74,0.77)	0.24 (0.23,0.26)	0.09 (0.06,0.13)	0.06 (0.02,0.11)	0.84 (0.80,0.89)
a2=0.1, c2=0.4, e2=0.5	0.00 (0.00,0.00)	0.82 (0.80,0.83)	0.18 (0.17,0.20)	0.09 (0.06,0.13)	0.24 (0.20,0.28)	0.66 (0.62,0.71)
a2=0.4, c2=0.2, e2=0.4	0.00 (0.00,0.00)	0.80 (0.79,0.81)	0.20 (0.19,0.21)	0.39 (0.36,0.43)	0.05 (0.00,0.09)	0.56 (0.52,0.60)
a2=0.4, c2=0.3, e2=0.3	0.00 (0.00,0.00)	0.83 (0.82,0.84)	0.17 (0.16,0.18)	0.39 (0.36,0.43)	0.14 (0.10,0.18)	0.47 (0.43,0.51)

6.2.2 Discussion

The main finding from this study is that the heritability of cause-specific survival among men with prostate cancer is approximately 10%, which is considerably lower compared with many other cancer traits[164] and in particular with prostate cancer susceptibility (58%)[55]. This supports the hypothesis that susceptibility of prostate cancer and progression of disease are separate mechanisms that involve different genes. Furthermore, the results of this study imply that the total additive genetic effect is small and that the predictive capacity in future prediction models may be low for this outcome. Another probable implication is that the amount of common variants with moderate effect sizes, usually found in GWAS, may be limited. However, many common variants with very small effect sizes or rare more penetrant variants may exist. Identifying these genetic variants associated with the aggressiveness and prognosis of prostate cancer is of great importance to enhance our understanding of aggressive prostate cancer etiology.

It is also important to emphasize that the ACE model, as used in this study, is rather crude since it relies on moderately strong assumptions. For example, the model assumes that no epistasis or gene-environment interactions exist. This issue has not been properly addressed for the studied outcome, since it requires much larger samples of men with follow-up data after prostate cancer diagnosis than are available today. Another assumption is that brothers are assumed to share environment, whereas fathers and sons do not. This may be a reasonable approximation for prostate cancer survival, but is not expected to be entirely accurate. However, this assumption is not empirically testable.

Publically available software implementations for analysis of family data with a survival outcome are currently rare or even nonexistent. Our conditional likelihood approach was implemented in the open source package OpenMX. We demonstrated that this application estimated the heritability well in simulated data. However, the shared environmental effect is expected to be under-estimated; the reason for this is not completely obvious. In the sensitivity analysis where T1c tumors (PSA detected without clinical symptoms of prostate cancer) were removed we saw that the estimate of the shared environment increased substantially. This may indicate that lead time bias, induced by opportunistic PSA screening, is the problem, although we adjust for this in the model. Thus, we cannot rule out the possibility of a common environmental effect, despite that it was estimated to be 0 in this study.

Furthermore, we considered other modeling strategies for this data. For example, we applied a Mixed Accelerated Failure Time (MAFT) model developed by Yip and coworkers[165]. This method maximizes a hierarchical likelihood procedure with a clever algorithm to speed up estimation. In an application on simulated data (with no covariate dependencies), we were not able to achieve satisfactory convergence for the model. This could be due to a poor model implementation or that the method makes too strong approximations in its algorithm. Another possibility is that the random effects for the simulated example are difficult to estimate, which has been shown for an ACE model with binary outcomes[166]. We encountered similar

problems when we applied a Bayesian method to the same simulated data[167]. Nevertheless, our conditional likelihood approach resulted in robust predictions in the simulation datasets, which provides some reassurance for robustness of our estimates of the heritability of prostate cancer-specific survival.

6.3 STUDY III: PROSTATE CANCER RISKVARIANTS ARE NOT ASSOCIATED WITH DISEASE PROGRESSION

6.3.1 Results

Results from association tests between 23 established risk SNPs with prostate cancer progression are shown in

Table *7*. Three of these genetic variants were nominally significant: rs9364554 (p=0.04) on chromosome 6q25 and rs10896449 (p= 0.03) on chromosome 11q13 among patients treated with curative intent (radiation therapy or radical prostatectomy); and rs4054823 (p=0.008) on chromosome 17p12 among patients on surveillance. However, none of these variants remained significant after Bonferroni correction. Furthermore, the cumulative effect of these SNPs was explored in a polygenic risk score but no significant association with prostate cancer progression was observed, neither for patients treated with curative intent (p=0.14) nor for those on surveillance (p=0.92).

6.3.2 Discussion

We did not find any of the 23 established prostate cancer susceptibility SNPs to be associated with prostate cancer progression among patients with localized disease at diagnosis. These results are in line with other studies that have addressed this issue[168,169]. However, some studies with similar cohorts have reported associations with known risk SNPs and disease prognosis outcomes, such as prostate cancer-specific mortality[102,170], aggressive prostate cancer[101] and biochemical recurrence[171-173]. Many of these studies are small and do not adjust for multiple testing (i.e. Bonferroni correction) in the reported associations. Furthermore, findings are inconclusive, with no overlap of observed associations between studies. The only exception from this are two SNPs (rs2735839 and rs17632542), located in the KLK3 gene (encoding PSA), which have been replicated in several studies[170,171,174]. We did not observe any association between rs2735839 and prostate cancer progression in our study. Shui and coworkers found that the susceptibility risk allele (G) for this SNP was inversely associated with prognosis[102]. Furthermore, it has been shown that the G-allele is associated with higher levels of circulating PSA[90,175,176]. Hence, individuals in an early stage of prostate cancer carrying this allele are more likely to be discovered by a PSA test as compared to individuals not carrying the allele. Thus, the protective effect on disease prognosis observed for this allele may be explained by PSA induced lead-time bias and not by a true biological association with prognosis.

Table 7: *Association between 23 established prostate cancer susceptibility variants and disease progression.*

		RADIATION+OPERATION		Surveillance	
SNP (CHR)	**Alleles***	**HR** (95% CI)**	**P-value*****	**HR** (95% CI)**	**P-value*****
rs721048 (2p15)	G/A	0.96(0.82-1.14)	0.66	0.94 (0.75–1.18)	0.86
rs2660753 (3p12)	C/T	1.03(0.83-1.28)	0.54	0.78(0.56-1.08)	0.26
rs629242 (4q12)	C/T	1.16(0.99-1.36)	0.19	1.22(0.99-1.51)	0.08
rs9364554 (6q25)	C/T	1.13(0.98-1.29)	0.04	1.06(0.88-1.27)	0.09
rs10486567 (7p15)	C/T	0.99(0.85-1.16)	0.74	0.90(0.72-1.11)	0.61
rs6465657(7q21)	C/T	1.10(0.97-1.25)	0.09	0.88(0.73-1.05)	0.34
rs16901979 (8q24)	C/A	0.83(0.61-1.13)	0.51	1.01(0.69-1.48)	0.35
rs6983267 (8q24)	G/T	1.08(0.95-1.23)	0.14	0.92(0.77-1.11)	0.42
rs1447295 (8q24)	C/A	1.07(0.89-1.27)	0.39	1.18(0.93-1.49)	0.21
rs1571801 (9q33)	G/T	0.96(0.83-1.11)	0.15	1.03(0.85-1.24)	0.76
rs10993994 (10q11)	C/T	0.94(0.82-1.07)	0.62	1.06(0.88-1.26)	0.74
rs10761581 (10q11)	T/G	1.08(0.95-1.23)	0.38	1.01(0.85-1.20)	0.98
rs4962416 (10q26)	A/G	1.06(0.91-1.24)	0.70	0.95(0.77-1.16)	0.83
rs12418451 (11q13)	G/A	1.08(0.93-1.24)	0.29	0.99(0.82-1.19)	0.89
rs10896449 (11q13)	G/A	0.89(0.78-1.02)	0.03	1.10(0.92-1.31)	0.37
rs4054823 (17p12)	T/C	0.98(0.86-1.12)	0.63	1.10(0.92-1.31)	0.008
rs11649743 (17q12)	C/T	1.16(0.98-1.37)	0.15	0.98(0.78-1.22)	0.93
rs4430796 (17q12)	T/C	1.05(0.92-1.21)	0.76	0.95(0.80-1.14)	0.79
rs1859962 (17q24)	G/T	0.92(0.80-1.05)	0.25	1.01(0.85-1.21)	0.83
rs887391 (19q13)	T/C	0.92(0.78-1.08)	0.12	0.96(0.78-1.18)	0.75
rs2735839 (19q13)	G/A	1.09(0.88-1.34)	0.56	1.30(1.00-1.68)	0.13
rs9623117 (22q13)	T/C	1.12(0.96-1.31)	0.35	0.95(0.76-1.17)	0.61
rs5945619 (Xp11)	A/G	1.03(0.85-1.24)	0.96	1.08(0.85-1.39)	0.82

* Common allele/Rare allele.
** Hazard ratio from additive Cox regression model.
*** Log-rank test p-value.

The main strengths of this study are the large sample size and the unique population-based design, complemented with follow-up regarding disease progression. The restricted follow-up time (mean 4 years, range 14 days to 8.5 years among patients without disease progression) may have limited the ability to explore the possible long-term effects of the assessed SNPs on prostate cancer progression. However, several studies report that disease progression after curative treatment is an early event among prostate cancer patients with a localized disease [177-180]. For example, Stephenson and coworkers report that the median time to biochemical recurrence was 22 months in a cohort of 3,125 men that were followed for 20 years[180]. The delayed collection of blood samples in the NPCR Follow-up study is another possible limitation. This may have resulted in a selection of patients carrying 'non-progressive' genetic variants that would bias our results towards null. However, in a vital status follow-up of the PROCAP participants in 2012, it was concluded that 134 patients (3.1%) of those that were included in the study in 2007 had died from prostate cancer[67]. Furthermore, Sullivan and coworkers report, from a similar cohort, that only 2% of the patients had died (from any cause) within 5 years after diagnosis[174]. Hence, we argue that death from prostate cancer is

rare among patients with a clinically localized disease and that the survival bias is of minor importance.

6.4 STUDY IV: GENOME-WIDE ASSOCIATION STUDY OF PROSTATE CANCER-SPECIFIC SURVIVAL

6.4.1 Results

In **Figure 9**, we can see the QQ plot of the test statistic from the combined prostate cancer-specific survival meta-analysis. The plot shows an early deviation of observed p-values from what is expected by a null distribution, which indicates possible inflation in the test statistic. However, the inflation factors, $\lambda = 1.11$ and $\lambda_{1000} = 1.02$ indicates no serious inflation in the test statistic.

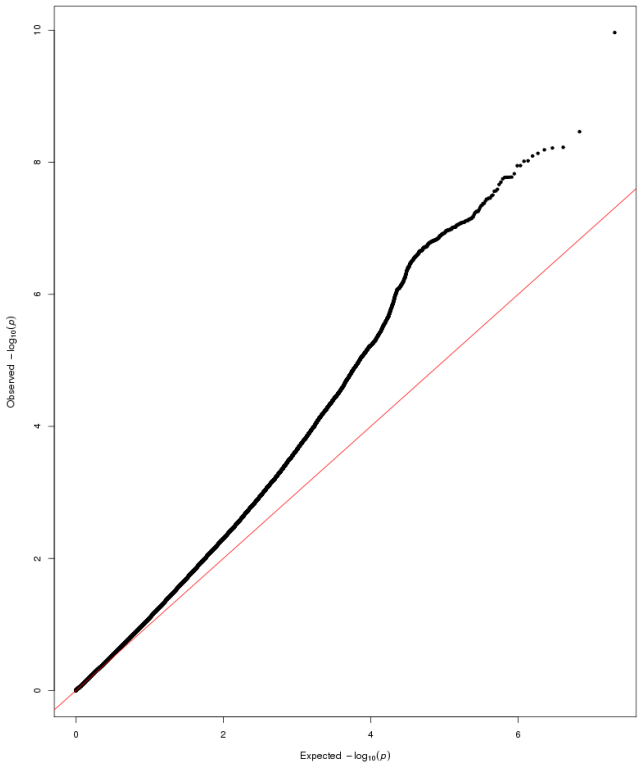

Figure 9: *QQ-plot of –log10(p) in combined meta-analysis between PRACTICAL and BPC3. The inflation factors were also assessed separately in PRACTICAL ($\lambda = 1.06$, $\lambda_{1000} = 1.02$) and BPC3 ($\lambda = 1.06$, $\lambda_{1000} = 1.06$).*

The Manhattan plot (**Figure 10**) of the combined meta-analysis shows in total 27 genome-wide significant SNPs located in chromosomes 1, 2, 3, 4, 5, 7, 8, 12, 18, 23. SNPs from 10 regions were selected for genotyping (based on the criteria, described in section **5.4.6**) to perform concordance analysis between imputed and genotyped data in the UKGPCS1 GWAS sample.

A summary of the 10 SNPs is shown in **Table 8**. Eight of these were rare variants with a minor allele frequency (MAF) of 1%-2% and two were common variants with MAF 7%–8%. Six of the SNPs failed genotyping in the UKGPCS1 sample, either because of failed assay design or monomorphism. The remaining four SNPs (rs114997855 on chromosome 2, rs76010824 on chromosome 3, rs140659849 and rs723557 on chromosome X) had an excellent concordance rate (98%–99%) between genotyped and imputed data. These were genotyped in the independent CONOR sample for replication. In CONOR, two SNPs (rs723557 and rs76010824) showed a null effect (HR=1) and two SNPs (rs114997855 and rs140659849) had opposite effects compared with the initial meta-analysis results. The combined effects of these four SNPs between PRACTICAL, BPC3 and CONOR were no longer genome-wide significant. Thus, none of the findings in the initial meta-analysis replicated.

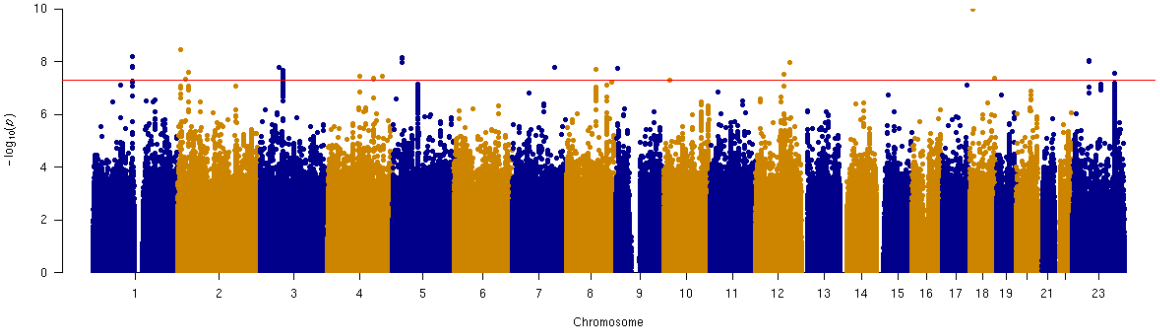

Figure 10: *Manhattan plot of p-values from prostate cancer-specific survival analysis. Red line marks genome-wide significant level, $p<5*10^{-8}$.*

6.4.2 Discussion

In this study we searched for SNPs associated with prostate cancer survival but we could not demonstrate any such association. This is in line with a smaller previous study with the same GWAS approach[108]. To this date, no genetic variant has been associated with this outcome on a genome-wide level ($p<5*10^{-8}$). However, five SNPs from a candidate gene study have been observed to associate with prostate cancer-specific survival[104] and were replicated in independent cohorts[105,106]. Furthermore, two SNPs on chromosome 3q26 and 5q14 were recently found to be associated with Gleason score in a case-only GWAS of prostate cancer[107]. Unfortunately we could only assess two of these seven variants in our study, which did not replicate.

In our initial meta-analysis between PRACTICAL and BPC3 we found several SNPs that were genome-wide significant, but none of them were replicated in the independent CONOR study. This underlines the importance of independent replication of GWAS findings from imputed data. Associations of imputed rare variants (representing the majority of our initial findings) are more uncertain in general. Furthermore, some of the SNPs were discovered in

smaller subsets of the full data due to poor imputation quality in many sub-cohorts, which increases the chance of random findings, such as false positives. Nevertheless, we were not able to follow-up six of the rare variants because of genotype failure. Thus, some of these

Table 8: *Summary of top findings in prostate cancer-specific survival analysis.*

		PRACTICAL and BPC3		CONOR	All studies[§]
SNP CHR:BP	**Alleles[ŧ] MAF**	**Total number No of PC/deaths**	**HR (95% CI) P-value**	**HR (95% CI) P-value**	**HR (95% CI) P-value**
rs190087062*	G/A	2,416/704	2.83 (1.99-4.02)		
1:115063785	0.02		6,5E-09		
rs114997855	A/G	20,051/2,729	1.75 (1.44-2.13)	0.88 (0.42-1.85)	1.67 (1.38-2.03)
2:30622824	0.02		2,6E-08	0,73	1,20E-07
rs76010824	A/G	23,251/3,324	1.29 (1.18-1.41)	1.01 (0.76-1.35)	1.26 (1.16-1.38)
3:67442642	0.07		2,8E-08	0,94	1,10E-07
rs184342703*	T/C	6,812/832	2.36 (1.73-3.20)		
4:135989066	0.02		4,2E-08		
rs192864713*	G/A	1,738/464	3.54 (2.31-5.43)		
5:27429220	0.01		7,3E-09		
rs111414857**	G/A	17,146/2,236	1.98 (1.56-2.50)		
7:126639415	0.01		1,7E-08		
rs149470135*	A/T	4,725/599	3.09 (2.09-4.59)		
8:86472701	0.01		2,0E-08		
rs117643112**	C/A	6,306/1,577	1.93 (1.53-2.43)		
12:81746712	0.02		3,1E-08		
rs140659849[ɋ]	A/G	2,702/271	3.00 (2.06-4.36)	0.75 (0.24-2.33)	2.61 (1.83-3.73)
X:50194937	0.01		9,6E-09	0,62	1,20E-07
rs723557[ɣ]	G/T	23,251/3,324	1.17 (1.10-1.24)	1.00 (0.84-1.19)	1.15 (1.09-1.22)
X:126653357	0.08		1,5E-07	0,98	6,10E-07

Abbreviations: CHR=Chromosome, BP=Base position (Genome build 37), MAF=Minor allele frequency, HR=Hazard ratio, PC=Prostate cancer, 95% CI=95% confidence interval.
ŧ Minor allele/Major allele. Minor allele used as effect allele (major as reference) in analysis.
§ Meta-analysis between PRACTICAL, BPC3 and CONOR.
ɣ Proxy for rs13440791 (p=2.7E-08 in PRACTICAL and BPC3).
ɋ Proxy for rs190977150 (p=9.5E-09 in PRACTICAL and BPC3).
* Monomorphic in UKGPCS1 replication.
** Failed assay (samples did not cluster) in UKGPCS1 replication.
*** Failed assay (failed quality control) in UKGPCS1 replication.

may have replicated if genotyping had succeeded, although it is rather unlikely given the weakness of the initial finding. However, it was surprising that the two more common variants, rs76010824 on chromosome 3 and rs723557 on chromosome X, which were identified in the full cohort, were not replicated. A possible explanation for rs723557 might be that it is a surrogate SNP for the initial GWAS finding (rs13440791), which we could not genotype because of manufacturing reasons. Although, this is not likely since these two SNPs were in perfect LD (r^2=1 in the European ancestry population of the phase 3 release of the 1000Genome project).

There is a possibility that population stratification affected our analysis, which was indicated by the QQ plot. However, since the adjusted inflation factor λ_{1000} (1.02) indicated minimal inflation, we chose not to include any principal components in the survival models to maintain statistical power. Participants in most studies were ascertained in a restricted geographical area (on sub-country basis, not nationwide). Thus, our model could be viewed as a geographically stratified analysis. Furthermore, we performed a number of sensitivity analyses to assess confounding and bias from different sources revealing robust results. Thus, we feel confident that the initial top findings were not confounded.

Another possible reason for the null finding in this study is that our analysis was based on imputed data and some areas of the genome were not well represented due to a low number of SNPs with good imputation quality, especially in studies genotyped with the iCOGS chip. Furthermore, we used a very heterogeneous group of patients, in terms of disease aggressiveness, local spread at diagnosis, mode of detection (screening or clinical symptoms) and treatment modalities. Moreover, several reasons why prostate cancer tumors develop to lethal disease exist, for example, unsuccessful radical prostatectomy, non-response to radiation therapy and resistance to hormonal treatment and initial chemotherapy. Each of these paths may involve different genes. Thus, analyzing all prostate cancer patients together may have diluted these effects. In this study we did not have the possibility to adjust or stratify for this properly due to insufficient information regarding diagnostic tumor characteristics and treatment regimens.

7 FUTURE DIRECTIONS

GWAS have identified more than 100 SNPs associated with prostate cancer incidence. Currently, these genetic variants explain approximately 39% of the familial risk[94], which indicates that more genes can be discovered. For example, it has been estimated that almost 2,000 additional SNPs that are truly associated with prostate cancer exist on the iCOGS chip[88]. Thus, larger consortia studies, such as the OncoArray initiative, are warranted to find these genetic variants. However, it is likely that risk SNPs that will be identified in future GWAS will have even smaller effect sizes than what we see today (OR<1.05). Thus, the individual contribution of these variants to prediction models will probably be small. Nevertheless, each small contribution will increase the predictive capacity. Furthermore, large-scale sequencing studies of the exome and eventually of the whole genome may reveal new rare prostate cancer risk variants with high or moderate penetrance (such as HOXB13 and the BRCA genes). Despite that rare variants explain a small part of prostate cancer heritability, they will enhance the understanding of prostate cancer etiology.

We have observed in this thesis that we can discriminate between prostate cancer cases and healthy controls, to some extent (AUC=0.68), based on the current knowledge of genetic risk variants (from the iCOGS chip). Furthermore, we saw that these SNPs may have benefits for risk stratification, and it has also recently been shown that over-diagnosis could be significantly reduced in a screening situation by using a polygenic risk score[181]. However, the predictive capacity of these SNPs alone is not yet of the magnitude that is required (AUC=0.75) to be beneficial in a screening context[182]. Nevertheless, as the knowledge of the biological function of these markers increase, predictions will also improve. A first step is to fine-map established prostate cancer risk regions to identify functional genes (or SNPs that are more correlated than current tagSNPs). Ideally, this would be performed with deep re-sequencing of these parts of the genome, coupled with functional annotation of regulatory elements and studies of gene expression in prostate cancer tumors. Amin Al Olama and coworkers found that the explained familial relative risk of prostate cancer increased with 9% when more functional variants were identified in their fine-mapping study[94]. Hence, by incorporating newly discovered SNPs and fine-mapping variants will probably increase predictions substantially.

Genetic variants have the potential to be effective biomarkers since they do not change over time, which is appealing since a blood test at birth could implicate the cancer risk later in life. However, it is likely that future prediction models will combine SNPs (including interactions) with other biomarkers and risk factors. There is evidence that such approaches are more successful. For example, it has been shown that predictions improved by combining SNPs with family history and PSA[95,183-185]. Furthermore, the STHLM1 study has reported that 22.7% of prostate biopsies could be avoided (at a cost of missing 3% diagnosis of patients characterized as having an aggressive disease) by adding a polygenic risk score with 35 SNPs to a model with age, family history, PSA and free-to-total PSA[186]. The utility of these combined models needs to be further evaluated, and translating the results to clinical practice

is a major task for future research. The STHLM3 study is an excellent example of a screening trial, where the purpose is to reduce the number of unnecessary biopsies and to maintain the sensitivity for aggressive tumors, based on 256 SNPs, a biomarker panel (Total PSA, Free PSA, Human kallikrein 2[187], Macrophage inhibitory cytokine-1[188] and microseminoprotein beta gene[189]) and clinical data. It is likely that similar models will be useful in a screening scenario in the near future.

In parallel with research that focus on discovery of new prostate cancer biomarkers, development of statistical methodology for prediction models that use SNP data is important. It is striking that the great majority of prediction models (including most prostate cancer prediction studies) are based on a simple regression model with a polygenic risk score. Genetic risk scores are popular because they have an intuitive polygenic additive interpretation that seems to fit data well. However, we know that undetected truly associated variants exist on current SNP platforms, which we were not able to identify with our methods. Thus, strategies that perform better variable selections are warranted. We applied more sophisticated prediction models (for example, a random forest model and various regularized regression methods), which according to the machine learning literature promise to perform this task better in high dimensional problems[130]. As previously mentioned, we could not see any improved predictions with these models. This could be due to the fact that SNPs are not well suited to these methods and that a polygenic risk score models the biology better. Nevertheless, assessment of existing machine learning methods in comparison with polygenic risk score models and methodological development in this field of research is needed.

The results from our heritability study suggest that genes may not be very useful as predictive markers for prostate cancer survival. However, our study was a first attempt to estimate this quantity and has to be re-confirmed in other study populations, preferably with other types of relatives (for example, twins with known zygosity, half-brothers or cousins). Our conditional likelihood approach is easily translated to other family structures that include pairs, but for more extended families other methods may be preferable. The MAFT model, which we tried to apply without much success, would be an appealing alternative since it easily extends to more complex family structures. The development of a publicly available implementation of this method is warranted. Furthermore, developments of methods that can adapt other transforms of the time outcome than the log-transform are needed. For example, one could use a Box-Cox transformation to approximate normality or use a bivariate Gumbel distribution applied in our conditional likelihood approach[190]. In our study this was not necessary since a log-normal distribution fit the data reasonably well, although this may not be true in general.

Genome-wide Complex Trait Analysis (GCTA) is a different approach to estimate the heritability of a trait. This method uses a genetic relationship matrix (estimated genetic correlations between individuals in a GWAS sample) to estimate the additive genetic part of the total variation in the outcome (i.e. narrow-sense heritability) by using a mixed linear

model[191]. One major benefit of GCTA is that no assumption regarding shared environment is needed, and unrelated individuals can be used. However, the method requires large population-based GWAS samples (in the order of 5,000 to 10,000 or more[192]), which for prostate cancer survival data is presently difficult to find, since most studies are enriched for cases with a high Gleason score, a family history of the disease or early onset disease. However, it might be possible to develop an application, as for binary outcomes, to adjust the method for ascertainment, to give population-based estimates[193]. Nevertheless, GCTA is not yet developed for survival outcomes but should in theory be possible to implement for such traits. This would give a nice independent confirmation of our results.

Despite that inherited germline SNPs may have low predictive capacity for prostate cancer survival, it is of great importance to identify genes associated with prostate cancer prognosis to enhance our understanding of aggressive prostate cancer etiology. Patients with a localized prostate cancer, as in the PROCAP study, are of particular interest since this group contains many indolent tumors, which today are hard to distinguish from early stage aggressive disease. We did not find any association between established risk variants and disease progression in this group of patients. Furthermore, accumulated evidence from several other studies points in the direction that GWAS SNPs associated with the development of prostate cancer are not involved in progression to aggressive lethal disease. However, it is important to search for SNPs associated with prostate cancer survival in other parts of the genome.

The results from our prostate cancer survival GWAS and heritability study indicate that common variants with moderate or large effect sizes on prostate cancer survival will be difficult to discover. The problem has been that in the early GWAS era much of the focus was on identification of SNPs associated with prostate cancer susceptibility. Thus, studies were not designed to find genetic variants that predict prognosis. However, deeper collaboration in PRACTICAL has emerged and the focus has shifted towards collecting follow-up data for the prostate cancer cases. This effort has resulted in a sample of approximately 58,000 prostate cancer patients (of which approximately 5,000 died from the disease) that have been genotyped on the OncoArray. Hopefully through this effort we will be able to identify SNPs that are genome-wide significantly associated with prostate cancer survival. Furthermore, designing large-scale sequencing studies to identify rare variants, associated with disease prognosis should also be a target for future research.

Cancer survival is determined by many factors, including the metastatic potential of tumors, treatment, response to treatment and early detection. It is likely that different genes affect these factors differently. Hopefully, GWAS with larger sample sizes will overcome this issue with heterogeneity in the outcome. Nevertheless, it would be useful to perform GWAS in homogenous groups of patients with more specific prognostic outcomes. For example, it would be interesting to study survival in different treatment groups or search for genes that are associated with resistance to treatment among patients on hormonal treatment. However, this is not feasible today because these studies do not achieve adequate statistical power, in particular the small number of deceased prostate cancer patients that are available would be

problematic in a sub-group analysis. Hence, to increase sample size, even larger consortia collaborations with good qualitative registration of clinical variables are essential. Pooling existing resources is the only way to develop the knowledge of prostate cancer genetics from GWAS in the future.

8 ACKNOWLEDGEMENTS

When summarizing this thesis I realize that I am extremely fortunate to have met so many smart and talented people during these years. So many people have been involved in my work and I am really grateful to everyone for that. Thank you all!

Fredrik Wiklund, my main supervisor who is probably the most positive guy I have ever met. I really appreciate that you are so calm and that you never see any problems, only possibilities. You have always been open to new ideas, always had time whenever I needed and been encouraging. Thank you! I never could have done this without your help!

Mark Clements, my co-supervisor. I am very grateful to have learned so much about programing, statistics and other "cool" things from you. I feel that my projects became much more fun when you got involved, although a little frustrating sometimes:-). It has been a lot of fun working with you!

Daphne Macris, my external mentor, who made the front cover to this thesis and always made sure that I had the best posters at conferences. Thank you! You are a great friend and I always appreciate your professional advice.

Robert Karlsson who was my roommate for many years at MEB. Thank you for all your help and for patiently answering all my stupid questions! Your skills in genetics and programing are very impressive!

Ralf Kuja-Halkola who introduced me to the world of structural equation modeling and OpenMX! You really saved my fourth project when I was stuck in another track. Thank you!

A special thanks to **Anna Johansson, Marie Reilly** and **Juni Palmgren,** who inspired me to start working at MEB.

Markus Aly, for generously sharing your clinical knowledge! I learned more about prostate cancer in one day with you in the clinic than I would if I spent months studying books.

Marie Jansson and **Camilla Ahlqvist** for guidance in the administrative jungle of KI. You made my life a lot simpler. Thank you!

Thank you MEB coworkers **Thomas Whitington, Martin Eklund, Mattias Rantalainen, Johan Lindberg, Henrik Grönberg, Daniel Klevebring** for collaboration and sharing your knowledge in prostate cancer and genetics.

It has been a great pleasure to be a part of the biostatistics group at MEB. Thank you for seminars, lunches, potlucks, all the fika moments and for creating such a friendly and professional atmosphere: **Flaminia Chiesa, Henric Olsson, Therese Andersson, Andreas Karlsson, Caroline Weibull, Sandra Eloranta, Cecilia Lundholm, Arvid Sjölander, Annika Tilliander, Linda Abrahamsson, Hatef Darabi, Andrea Ganna, Peter Ström, Henric Winell, Xingrong Li, Johan Zetterqvist, Hannah Bower, Gabriel Isheden,**

Elisabeth Dahlqwist, Bénédicte Delcoigne, Keith Humphreys, Alexander Ploner, Yudi Pawitan, Paul Dickman and **Paul Lambert.**

Other MEBers that I would like to thank: **Elisabeth Möller, Stephanie Bonn, Adina Feldman, Thomas Frisell** and **Patrik Magnusson.**

I am also very grateful to all members of the PRACTICAL consortium. Without your data collection efforts and generosity to share data, two of my projects would not have been possible to accomplish. In particular I would like to thank **Rosalind Eeeles, Sara Benlloch, Ali Amin Al Olama, Douglas Easton, Koveela Govindasami** and **Zsofia Kote-Jarai.**

Jan Sunquist, **Lars Agréus** and **Gunnar Nilsson** for being supportive of my PhD studies, and making it possible to work in parallel at former Centrum för allmänmedicin (CeFAM). I would also like to acknowledge **Sven-Erik Johansson** at CeFAM, you have always been a role model for me.

Big thanks to my dear friends **Ulf Eriksson, Emanuel Mörk, Rasmus Sundin, Sami Soliman** and **Georges Mansourati**. You all deserve a page of acknowledgments for your friendship and support. Let's take it over a beer instead. Really looking forward to that!

Special thanks to my whole family! **Dad**, it has always been inspiring to see your dedication and love to science. This is probably one of the reasons why I am where I am. **Mom**, thank you for always believing in me and for being the best grandmother to my children. **Marta**, thank you for being the best sister and aunt. **Janne, Lena and Johanna,** thank you for all your help and support (I would not have been able to finish this thesis without your help). **Adam**, thanks for the support, especially in the end of this thesis! Thank you to the rest of my family: **Jaś, Rysiek, Anna, Olek, Terés, Cornelia, Julian, Magda, Johan, Maciek, Dominika, Ela, Tadeusz, Babcia, Berni, Anthony, Marysia and Tadeusz (mały).**

I am extremely proud of being a father of two such wonderful children as **Alma** and **Elliot**. Coming home to you, after sometimes having a hard day at work, has been fantastic. Elliot, I enjoy every time when you come running to me, as soon as I get inside the door, to tell me something important about your day (about how you went with the blue bus, ate pancakes, jumped a trampoline or something else important). Alma, there is nothing as wonderful as seeing your fantastic smile. Thank you both for reminding me every day about what is important in life!

To my wife **Emma**, I don't really know how I could ever thank you? You have been a greater support than I could ever wish for during this whole PhD journey. Without you, nothing in my life would have been half as good as it is. You are the best thing that ever happened to me! Thank you for being the love of my life!

9 REFERENCES

1. Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet].
2. Segi, M. Cancer Mortality for Selected Sites in 24 Countries (1950–57). Department of Public Health, Tohoku University of Medicine, Sendai, Japan. (1960).
3. McNeal, J.E. Regional morphology and pathology of the prostate. *American journal of clinical pathology* **49**, 347-357 (1968).
4. McNeal, J.E., Redwine, E.A., Freiha, F.S. & Stamey, T.A. Zonal distribution of prostatic adenocarcinoma. Correlation with histologic pattern and direction of spread. *The American journal of surgical pathology* **12**, 897-906 (1988).
5. Prins, G.S. What is the prostate and what are its functions? in *Handbook of Andrology* (ed. Robaire, B., Chan, P.) (The American Society of Andrology, 2010).
6. https://commons.wikimedia.org/wiki/File:Male_anatomy.png.
7. Berry, S.J., Coffey, D.S., Walsh, P.C. & Ewing, L.L. The development of human benign prostatic hyperplasia with age. *J Urol* **132**, 474-479 (1984).
8. Elkahwaji, J.E. The role of inflammatory mediators in the development of prostatic hyperplasia and prostate cancer. *Research and reports in urology* **5**, 1-10 (2012).
9. McVary, K.T. BPH: epidemiology and comorbidities. *The American journal of managed care* **12**, S122-128 (2006).
10. Chute, C.G., *et al.* The prevalence of prostatism: a population-based survey of urinary symptoms. *J Urol* **150**, 85-89 (1993).
11. McNeal, J.E. Normal histology of the prostate. *The American journal of surgical pathology* **12**, 619-633 (1988).
12. McNeal, J.E. Normal anatomy of the prostate and changes in benign prostatic hypertrophy and carcinoma. *Seminars in ultrasound, CT, and MR* **9**, 329-334 (1988).
13. De Nunzio, C., *et al.* The controversial relationship between benign prostatic hyperplasia and prostate cancer: the role of inflammation. *European urology* **60**, 106-117 (2011).
14. Lummus, W.E. & Thompson, I. Prostatitis. *Emergency medicine clinics of North America* **19**, 691-707 (2001).
15. Leigh, D.A. Prostatitis--an increasing clinical problem for diagnosis and management. *The Journal of antimicrobial chemotherapy* **32 Suppl A**, 1-9 (1993).
16. Stewart, C. Prostatitis. *Emergency medicine clinics of North America* **6**, 391-402 (1988).
17. Jiang, J., *et al.* The role of prostatitis in prostate cancer: meta-analysis. *PloS one* **8**, e85179 (2013).
18. Roberts, R.O., Bergstralh, E.J., Bass, S.E., Lieber, M.M. & Jacobsen, S.J. Prostatitis as a risk factor for prostate cancer. *Epidemiology* **15**, 93-99 (2004).

19. Nakai, Y. & Nonomura, N. Inflammation and prostate carcinogenesis. *International journal of urology : official journal of the Japanese Urological Association* **20**, 150-160 (2013).

20. Strasner, A. & Karin, M. Immune Infiltration and Prostate Cancer. *Frontiers in oncology* **5**, 128 (2015).

21. Sfanos, K.S. & De Marzo, A.M. Prostate cancer and inflammation: the evidence. *Histopathology* **60**, 199-215 (2012).

22. Sfanos, K.S., Isaacs, W.B. & De Marzo, A.M. Infections and inflammation in prostate cancer. *American journal of clinical and experimental urology* **1**, 3-11 (2013).

23. Powell, I.J., Bock, C.H., Ruterbusch, J.J. & Sakr, W. Evidence supports a faster growth rate and/or earlier transformation to clinically significant prostate cancer in black than in white American men, and influences racial progression and mortality disparity. *J Urol* **183**, 1792-1796 (2010).

24. Bell, K.J., Del Mar, C., Wright, G., Dickinson, J. & Glasziou, P. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. *International journal of cancer. Journal international du cancer* **137**, 1749-1757 (2015).

25. Arora, R., *et al.* Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate. *Cancer* **100**, 2362-2366 (2004).

26. Hughes, C., Murphy, A., Martin, C., Sheils, O. & O'Leary, J. Molecular pathology of prostate cancer. *Journal of clinical pathology* **58**, 673-684 (2005).

27. Lindberg, J., *et al.* Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. *European urology* **63**, 347-353 (2013).

28. Cooper, C.S., *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet* (2015).

29. http://www.internetmedicin.se/page.aspx?id=606.

30. Cohen, R.J., *et al.* Central zone carcinoma of the prostate gland: a distinct tumor type with poor prognostic features. *J Urol* **179**, 1762-1767; discussion 1767 (2008).

31. Grignon, D.J. & Sakr, W.A. Zonal origin of prostatic adenocarcinoma: are there biologic differences between transition zone and peripheral zone adenocarcinomas of the prostate gland? *Journal of cellular biochemistry. Supplement* **19**, 267-269 (1994).

32. Noguchi, M., Stamey, T.A., Neal, J.E. & Yemoto, C.E. An analysis of 148 consecutive transition zone cancers: clinical and histological characteristics. *J Urol* **163**, 1751-1755 (2000).

33. Augustin, H., *et al.* Zonal location of prostate cancer: significance for disease-free survival after radical prostatectomy? *Urology* **62**, 79-85 (2003).

34. Lindberg, J., Kristiansen, A., Wiklund, P., Gronberg, H. & Egevad, L. Tracking the Origin of Metastatic Prostate Cancer. *European urology* (2014).

35. Haffner, M.C., *et al.* Tracking the clonal origin of lethal prostate cancer. *The Journal of clinical investigation* **123**, 4918-4922 (2013).

36. Heidenreich, A., *et al.* EAU guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent-update 2013. *European urology* **65**, 124-137 (2014).

37. Heidenreich, A., *et al.* EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer. *European urology* **65**, 467-479 (2014).

38. Bratt, O., Damber, L-E.,Aldehag, A., Brändström, H. Prostatacancer – nya riktlinjer och första vårdprogrammet. *Läkartidinigen* **05/2015**(2015).

39. Gleason, D.F. Classification of prostatic carcinomas. *Cancer chemotherapy reports. Part 1* **50**, 125-128 (1966).

40. Epstein, J.I., Allsbrook, W.C., Jr., Amin, M.B., Egevad, L.L. & Committee, I.G. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *The American journal of surgical pathology* **29**, 1228-1242 (2005).

41. *American Joint Committee on Cancer (AJCC) Cancer Staging Manual, 7th Edition.*, (2010).

42. https://cancerstaging.org/references-tools/quickreferences/Documents/ProstateSmall.pdf.

43. Van der Kwast, T.H. & Roobol, M.J. Defining the threshold for significant versus insignificant prostate cancer. *Nature reviews. Urology* **10**, 473-482 (2013).

44. Holmstrom, B., *et al.* Prostate specific antigen for early detection of prostate cancer: longitudinal study. *Bmj* **339**, b3537 (2009).

45. Gann, P.H., Hennekens, C.H. & Stampfer, M.J. A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. *JAMA* **273**, 289-294 (1995).

46. Wolf, A.M., *et al.* American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA: a cancer journal for clinicians* **60**, 70-98 (2010).

47. Andriole, G.L., *et al.* Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. *J Natl Cancer Inst* **104**, 125-132 (2012).

48. Schroder, F.H., *et al.* Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet* **384**, 2027-2035 (2014).

49. Nordstrom, T., *et al.* Prostate-specific antigen (PSA) testing is prevalent and increasing in Stockholm County, Sweden, Despite no recommendations for PSA screening: results from a population-based study, 2003-2011. *European urology* **63**, 419-425 (2013).

50. http://www-dep.iarc.fr/NORDCAN/SW/frame.asp.

51. World Health Organization, mortality database http://www.who.int/healthinfo/statistics/mortality_rawdata/en/index.html (accessed on 28/05/2015).

52. Ferlay, J., Bray, F., Steliarova-Foucher, E., Forman, D. Cancer Incidence in Five Continents, CI5plus. IARC CancerBase No. 9. (2014).

53. Johns, L.E. & Houlston, R.S. A systematic review and meta-analysis of familial prostate cancer risk. *BJU international* **91**, 789-794 (2003).

54. Brandt, A., Bermejo, J.L., Sundquist, J. & Hemminki, K. Age-specific risk of incident prostate cancer and risk of death from prostate cancer defined by the number of affected family members. *European urology* **58**, 275-280 (2010).

55. Hjelmborg, J.B., *et al.* The heritability of prostate cancer in the Nordic Twin Study of Cancer. *Cancer Epidemiol Biomarkers Prev* **23**, 2303-2310 (2014).

56. Lindstrom, L.S., *et al.* Familial concordance in cancer survival: a Swedish population-based study. *Lancet Oncol* **8**, 1001-1006 (2007).

57. Hemminki, K., Ji, J., Forsti, A., Sundquist, J. & Lenner, P. Concordance of survival in family members with prostate cancer. *J Clin Oncol* **26**, 1705-1709 (2008).

58. Hemminki, K. Familial risk and familial survival in prostate cancer. *World journal of urology* **30**, 143-148 (2012).

59. Morton, R.A., Jr. Racial differences in adenocarcinoma of the prostate in North American men. *Urology* **44**, 637-645 (1994).

60. Taioli, E., *et al.* Multi-institutional prostate cancer study of genetic susceptibility in populations of African descent. *Carcinogenesis* **32**, 1361-1365 (2011).

61. Gronberg, H. Prostate cancer epidemiology. *Lancet* **361**, 859-864 (2003).

62. Shimizu, H., *et al.* Cancers of the prostate and breast among Japanese and white immigrants in Los Angeles County. *British journal of cancer* **63**, 963-966 (1991).

63. Cook, L.S., Goldoft, M., Schwartz, S.M. & Weiss, N.S. Incidence of adenocarcinoma of the prostate in Asian immigrants to the United States and their descendants. *J Urol* **161**, 152-155 (1999).

64. Giovannucci, E., Liu, Y., Platz, E.A., Stampfer, M.J. & Willett, W.C. Risk factors for prostate cancer incidence and progression in the health professionals follow-up study. *International journal of cancer. Journal international du cancer* **121**, 1571-1578 (2007).

65. Chan, J.M., Gann, P.H. & Giovannucci, E.L. Role of diet in prostate cancer development and progression. *J Clin Oncol* **23**, 8152-8160 (2005).

66. Hackshaw-McGeagh, L.E., *et al.* A systematic review of dietary, nutritional, and physical activity interventions for the prevention of prostate cancer progression and mortality. *Cancer Causes Control* (2015).

67. Bonn, S.E., *et al.* Body mass index and weight change in men with prostate cancer: progression and mortality. *Cancer Causes Control* **25**, 933-943 (2014).

68. Cao, Y. & Ma, J. Body mass index, prostate cancer-specific mortality, and biochemical recurrence: a systematic review and meta-analysis. *Cancer prevention research* **4**, 486-501 (2011).

69. Davies, N.M., *et al.* The effects of height and BMI on prostate cancer incidence and mortality: a Mendelian randomization study in 20,848 cases and 20,214 controls from the PRACTICAL consortium. *Cancer Causes Control* (2015).

70. Thomas, D. *Statistical Methods In Genetic Epidemiology*, (2004).

71. https://commons.wikimedia.org/wiki/File:Dna-SNP.svg.

72. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

73. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

74. Ng, S.B., *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).

75. Consortium, E.P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, e1001046 (2011).

76. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-796 (2003).

77. http://www.uk10k.org/.

78. Genomes Project, C., *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).

79. Kote-Jarai, Z., *et al.* BRCA2 is a moderate penetrance gene contributing to young-onset prostate cancer: implications for genetic testing in prostate cancer patients. *British journal of cancer* **105**, 1230-1234 (2011).

80. Castro, E. & Eeles, R. The role of BRCA1 and BRCA2 in prostate cancer. *Asian journal of andrology* **14**, 409-414 (2012).

81. Edwards, S.M., *et al.* Prostate cancer in BRCA2 germline mutation carriers is associated with poorer prognosis. *British journal of cancer* **103**, 918-924 (2010).

82. Tryggvadottir, L., *et al.* Prostate cancer progression and survival in BRCA2 mutation carriers. *J Natl Cancer Inst* **99**, 929-935 (2007).

83. Leongamornlert, D., *et al.* Germline BRCA1 mutations increase prostate cancer risk. *British journal of cancer* **106**, 1697-1701 (2012).

84. Karlsson, R., *et al.* A population-based assessment of germline HOXB13 G84E mutation and prostate cancer risk. *European urology* **65**, 169-176 (2014).

85. Huang, Q., *et al.* A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat Genet* **46**, 126-135 (2014).

86. Al Olama, A.A., *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103-1109 (2014).

87. Al Olama, A.A., *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41**, 1058-1060 (2009).

88. Eeles, R.A., *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* **45**, 385-391, 391e381-382 (2013).

89. Eeles, R.A., *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-1121 (2009).

90. Eeles, R.A., *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40**, 316-321 (2008).

91. Amundadottir, L.T., *et al.* A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**, 652-658 (2006).

92. Gudmundsson, J., *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* **40**, 281-283 (2008).

93. Gudmundsson, J., *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-637 (2007).

94. Amin Al Olama, A., *et al.* Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. *Hum Mol Genet* (2015).

95. Zheng, S.L., *et al.* Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* **358**, 910-919 (2008).

96. Hazelett, D.J., *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet* **10**, e1004102 (2014).

97. Penney, K.L., *et al.* Association of prostate cancer risk variants with gene expression in normal and tumor tissue. *Cancer Epidemiol Biomarkers Prev* **24**, 255-260 (2015).

98. Kote-Jarai, Z., *et al.* Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. *Human genetics* **129**, 687-694 (2011).

99. Kote-Jarai, Z., *et al.* Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression. *Hum Mol Genet* **22**, 2520-2528 (2013).

100. Goh, C.L., *et al.* Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *Journal of internal medicine* **271**, 353-365 (2012).

101. Amin Al Olama, A., *et al.* A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. *Hum Mol Genet* **22**, 408-415 (2013).

102. Shui, I.M., *et al.* Prostate cancer (PCa) risk variants and risk of fatal PCa in the National Cancer Institute Breast and Prostate Cancer Cohort Consortium. *European urology* **65**, 1069-1075 (2014).

103. Helfand, B.T., *et al.* Associations of prostate cancer risk variants with disease aggressiveness: results of the NCI-SPORE Genetics Working Group analysis of 18,343 cases. *Human genetics* **134**, 439-450 (2015).

104. Lin, D.W., *et al.* Genetic variants in the LEPR, CRY1, RNASEL, IL4, and ARVCF genes are prognostic markers of prostate cancer-specific mortality. *Cancer Epidemiol Biomarkers Prev* **20**, 1928-1936 (2011).

105. Karyadi, D.M., *et al.* Confirmation of genetic variants associated with lethal prostate cancer in a cohort of men from hereditary prostate cancer families. *International journal of cancer. Journal international du cancer* **136**, 2166-2171 (2015).

106. Penney, K.L., *et al.* Replication of a genetic variant for prostate cancer-specific mortality. *Prostate cancer and prostatic diseases* **18**, 260-263 (2015).

107. Berndt, S.I., *et al.* Two susceptibility loci identified for prostate cancer aggressiveness. *Nature communications* **6**, 6889 (2015).

108. Penney, K.L., *et al.* Genome-wide association study of prostate cancer mortality. *Cancer Epidemiol Biomarkers Prev* **19**, 2869-2876 (2010).

109. Lindstrom, S., *et al.* Germ-line genetic variation in the key androgen-regulating genes androgen receptor, cytochrome P450, and steroid-5-alpha-reductase type 2 is important for prostate cancer development. *Cancer Res* **66**, 11077-11083 (2006).

110. http://ki.se/en/meb/cancer-of-the-prostate-in-sweden-caps.

111. Liu, W., *et al.* Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. *Cancer Res* **69**, 2176-2179 (2009).

112. http://www.icr.ac.uk/our-research/research-divisions/division-of-genetics-and-epidemiology/oncogenetics/research-projects/ukgpcs.

113. Schumacher, F.R., *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet* **20**, 3867-3875 (2011).

114. http://epi.grants.cancer.gov/BPC3/.

115. http://practical.ccge.medschl.cam.ac.uk/.

116. http://www.cgems.cancer.gov/.

117. http://ccge.medschl.cam.ac.uk/files/2014/03/iCOGS_detailed_lists_ALL1.pdf.

118. Naess, O., *et al.* Cohort profile: cohort of Norway (CONOR). *International journal of epidemiology* **37**, 481-485 (2008).

119. http://www.socialstyrelsen.se/register/halsodataregister/cancerregistret/inenglish.

120. Barlow, L., Westergren, K., Holmberg, L. & Talback, M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. *Acta Oncol* **48**, 27-33 (2009).

121. Ekbom, A. The Swedish Multi-generation Register. *Methods in molecular biology* **675**, 215-220 (2011).

122. https://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/18019/2010-4-33.pdf.

123. Adolfsson, J., *et al.* Clinical characteristics and primary treatment of prostate cancer in Sweden between 1996 and 2005. *Scand J Urol Nephrol* **41**, 456-477 (2007).

124. Van Hemelrijck, M., *et al.* Cohort Profile: the National Prostate Cancer Register of Sweden and Prostate Cancer data Base Sweden 2.0. *International journal of epidemiology* **42**, 956-967 (2013).

125. Tomic, K., *et al.* Evaluation of data quality in the National Prostate Cancer Register of Sweden. *European journal of cancer* **51**, 101-111 (2015).

126. Tomic, K., *et al.* Capture rate and representativity of The National Prostate Cancer Register of Sweden. *Acta Oncol* **54**, 158-163 (2015).

127. Stattin, P., *et al.* Outcomes in localized prostate cancer: National Prostate Cancer Register of Sweden follow-up study. *J Natl Cancer Inst* **102**, 950-958 (2010).

128. Stattin, P., *et al.* Surveillance and deferred treatment for localized prostate cancer. Population based study in the National Prostate Cancer Register of Sweden. *J Urol* **180**, 2423-2429; discussion 2429-2430 (2008).

129. Guyon, I., Elisseeff, A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* **3**, 1157-1182 (2003).

130. Hastie, T., Tibshirani, R., Friedman, J. *The elements of statistical learning. Data mining, inference, and prediction*, (2009).

131. International Schizophrenia, C., *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).

132. Hanley, J.A. & McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36 (1982).

133. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* **6**, e1000864 (2010).

134. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).

135. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387-415 (1975).

136. Pepe, M.S., Janes, H., Longton, G., Leisenring, W. & Newcomb, P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology* **159**, 882-890 (2004).

137. Ware, J.H. The limitations of risk factors as prognostic tools. *N Engl J Med* **355**, 2615-2617 (2006).

138. Pencina, M.J., D'Agostino, R.B., Sr., D'Agostino, R.B., Jr. & Vasan, R.S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine* **27**, 157-172; discussion 207-112 (2008).

139. Cox, D.R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society* **34**, 187-220 (1972).

140. Cox, D.R. Partial likelihood. *Biometrika* **62**, 269-276 (1975).

141. Pawitan, Y. *In All Likelihood: Statistical modelling and Inference Using Likelihood*, (Oxford University Press, 2001).

142. Therneau, T.M. & Li, H. Computing the Cox model for case cohort designs. *Lifetime data analysis* **5**, 99-112 (1999).

143. Aalen, O.O., Borgan, O., Gjessing, H. K. *Survival and Event History Analysis. A process point of view*, (Springer, 2008).

144. Plomin, R. *Behavioral genetics*, (Worth Publisher, 2013).

145. Wright, S. Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics* **6**, 111-123 (1921).

146. Rijsdijk, F.V. & Sham, P.C. Analytic approaches to twin data using structural equation models. *Briefings in bioinformatics* **3**, 119-133 (2002).

147. Verweij, K.J., Mosing, M.A., Zietsch, B.P. & Medland, S.E. Estimating heritability from twin studies. *Methods in molecular biology* **850**, 151-170 (2012).

148. http://www.statistikdatabasen.scb.se/.

149. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics* **9**, 477-485 (2008).

150. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-1296 (2007).

151. Novembre, J., *et al.* Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).

152. Humphreys, K., *et al.* The genetic structure of the Swedish population. *PloS one* **6**, e22547 (2011).

153. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

154. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11**, 499-511 (2010).

155. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annual review of genomics and human genetics* **10**, 387-406 (2009).

156. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-959 (2012).

157. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* **34**, 816-834 (2010).

158. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010).

159. Bulik-Sullivan, B.K., *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).

160. de Bakker, P.I., *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-128 (2008).

161. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).

162. Reich, D.E. & Goldstein, D.B. Detecting association in a case-control study while correcting for population stratification. *Genetic epidemiology* **20**, 4-16 (2001).

163. Amin Al Olama, A., *et al.* Risk Analysis of Prostate Cancer in PRACTICAL, a Multinational Consortium, Using 25 Known Prostate Cancer Susceptibility Loci. *Cancer Epidemiol Biomarkers Prev* **24**, 1121-1129 (2015).

164. Lichtenstein, P., *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85 (2000).

165. Yip, B.H., Moger, T.A. & Pawitan, Y. Genetic analysis of age-at-onset traits based on case-control family data. *Statistics in medicine* **29**, 3258-3266 (2010).

166. Kuhnert, P.M. & Do, K.A. Fitting genetic models to twin data with binary and ordered categorical responses: a comparison of structural equation modelling and Bayesian hierarchical models. *Behavior genetics* **33**, 441-454 (2003).

167. Scurrah, K.J., Palmer, L.J. & Burton, P.R. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genetic epidemiology* **19**, 127-148 (2000).

168. Penney, K.L., *et al.* Evaluation of 8q24 and 17q risk loci and prostate cancer mortality. *Clin Cancer Res* **15**, 3223-3230 (2009).

169. Wiklund, F.E., *et al.* Established prostate cancer susceptibility variants are not associated with disease outcome. *Cancer Epidemiol Biomarkers Prev* **18**, 1659-1662 (2009).

170. Pomerantz, M.M., *et al.* Association of prostate cancer risk Loci with disease aggressiveness and prostate cancer-specific mortality. *Cancer prevention research* **4**, 719-728 (2011).

171. Gallagher, D.J., *et al.* Susceptibility loci associated with prostate cancer progression and mortality. *Clin Cancer Res* **16**, 2819-2832 (2010).

172. Huang, S.P., *et al.* Prognostic significance of prostate cancer susceptibility variants on prostate-specific antigen recurrence after radical prostatectomy. *Cancer Epidemiol Biomarkers Prev* **18**, 3068-3074 (2009).

173. Cheng, I., *et al.* Prostate Cancer Susceptibility Variants Confer Increased Risk of Disease Progression. *Cancer Epidemiol Biomarkers Prev* (2010).

174. Sullivan, J., *et al.* An analysis of the association between prostate cancer risk loci, PSA levels, disease aggressiveness and disease-specific mortality. *British journal of cancer* **113**, 166-172 (2015).

175. Ahn, J., *et al.* Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. *Nat Genet* **40**, 1032-1034; author reply 1035-1036 (2008).

176. Gudmundsson, J., *et al.* Genetic correction of PSA values using sequence variants associated with PSA levels. *Science translational medicine* **2**, 62ra92 (2010).

177. Taira, A.V., *et al.* Time to failure after definitive therapy for prostate cancer: implications for importance of aggressive local treatment. *Journal of contemporary brachytherapy* **5**, 215-221 (2013).

178. Kollmeier, M.A., Stock, R.G. & Stone, N. Biochemical outcomes after prostate brachytherapy with 5-year minimal follow-up: importance of patient selection and implant quality. *International journal of radiation oncology, biology, physics* **57**, 645-653 (2003).

179. Uchio, E.M., Aslan, M., Wells, C.K., Calderone, J. & Concato, J. Impact of biochemical recurrence in prostate cancer among US veterans. *Archives of internal medicine* **170**, 1390-1395 (2010).

180. Stephenson, A.J., *et al.* Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. *J Clin Oncol* **24**, 3973-3978 (2006).

181. Pashayan, N., *et al.* Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genetics in medicine : official journal of the American College of Medical Genetics* (2015).

182. Janssens, A.C., *et al.* The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genetics in medicine : official journal of the American College of Medical Genetics* **9**, 528-535 (2007).

183. Lindstrom, S., *et al.* Common genetic variants in prostate cancer risk prediction--results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). *Cancer Epidemiol Biomarkers Prev* **21**, 437-444 (2012).

184. Chatterjee, N., Wheeler, B., et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics* **45**(2013).

185. So, H.C., Kwan, J.S., Cherny, S.S. & Sham, P.C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *American journal of human genetics* **88**, 548-565 (2011).

186. Aly, M., *et al.* Polygenic risk score improves prostate cancer risk prediction: results from the Stockholm-1 cohort study. *European urology* **60**, 21-28 (2011).

187. Schedlich, L.J., Bennetts, B.H. & Morris, B.J. Primary structure of a human glandular kallikrein gene. *DNA* **6**, 429-437 (1987).

188. Husaini, Y., *et al.* Macrophage inhibitory cytokine-1 (MIC-1/GDF15) slows cancer development but increases metastases in TRAMP prostate cancer prone mice. *PloS one* **7**, e43833 (2012).

189. Chang, B.L., *et al.* Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Hum Mol Genet* **18**, 1368-1375 (2009).

190. Demirhan, H., Kalaylioglu, Z. On the generalized multivariate Gumbel distribution. *Statistics & Probability Letters* **103**, 93-99 (2015).

191. Yang, J., *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569 (2010).

192. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Human genetics* **131**, 1655-1664 (2012).

193. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *American journal of human genetics* **88**, 294-305 (2011).