GENETIC VARIATION
AND PROSTATE CANCER
Population-based association studies in Sweden

SARA LINDSTRÖM

Umeå 2007

From the Department of Radiation Sciences, Oncology
Umeå University, Umeå, Sweden
To my parents
Abstract

Prostate cancer constitutes the most common malignancy and the most common cause of cancer-related death in Swedish men. A large body of evidence suggests that inherited genetic variants contribute to both development and progression of prostate cancer. The aim of this thesis is to identify genetic variants that alter prostate cancer risk and progression. All papers included in this thesis are based on a Swedish population-based case-control study (CAPS) comprising 2,965 incident prostate cancer cases and 1,823 controls.

In paper I, we investigated if genetic variants in the E-cadherin gene altered prostate cancer risk. Seven haplotype tagging SNPs (tagSNPs) were selected and genotyped in CAPS and families with hereditary prostate cancer. We confirmed association of a promoter SNP rs16260 previously reported to increase risk of hereditary prostate cancer (OR: 2.6; 95% CI: 1.6-4.3) for homozygous ‘A’ carriers.

In paper II, we assessed 46 polymorphisms earlier reported to be associated with prostate cancer risk. Six polymorphisms in five different genes were replicated. Interestingly, three of these genes were involved in the androgen biosynthesis.

In paper III, we followed up on the results from paper II by genotyping 23 tagSNPs located in the hormone regulating genes AR, CYP17 and SRD5A2. Multiple SNPs and haplotypes were associated with prostate cancer risk, especially in the AR gene. Combining risk alleles from all genes revealed a substantial risk increase for each additional allele carried (OR: 1.12; 95% CI: 1.1-1.2, P=0.00009).

In paper IV, we collected information about cause of death for all case patients in CAPS. At time of follow-up 300 study subjects were deceased from prostate cancer. We assessed AR, CYP17 and SRD5A2 variants for association with lethal prostate cancer and found overall no association. However, one AR promoter SNP was associated with an increased risk of dying from prostate cancer amongst men who received palliative hormonal therapy as primary treatment.

In paper V, we assessed common genetic variation at the ERG locus for association between prostate cancer risk and survival. ERG is recognized as a proto-oncogene frequently overexpressed in prostate cancer. A total of 21 tagSNPs in the 5’ region of ERG were genotyped. There was no correlation between ERG SNPs and prostate cancer risk but common genetic variation located approximately 100,000
basepairs upstream of ERG was significantly associated with prostate cancer-specific survival.

In summary, our results suggest that common genetic variation in E-cadherin alters prostate cancer risk in Swedish men with a positive family history of prostate cancer. Moreover, common genetic variation in the androgen-related genes AR, CYP17 and SRD5A2 affects the risk of developing prostate cancer but is unlikely to alter prostate cancer progression. However, genetic variants in AR may affect hormonal therapy response. Finally, ERG polymorphisms are associated with prostate cancer-specific death but are not likely to play a role in prostate cancer development.
Contents

LIST OF ORIGINAL PUBLICATIONS ................................................................................................................. 8

1. BACKGROUND ................................................................................................................................................. 9

1.1 GENETIC VARIATION ................................................................................................................................. 9
1.1.1 OUR HUMAN GENOME AND ITS VARIABILITY ....................................................................................... 9
1.1.2 SINGLE NUCLEOTIDE POLYMORPHISMS .......................................................................................... 10
1.1.3 LINKAGE DISEQUILIBRIUM .................................................................................................................... 10
1.1.4 HAPLOTYPES AND HAPLOTYPE BLOCKS ............................................................................................. 11
1.1.5 HAPLOTYPE TAGGING ............................................................................................................................ 12

1.2 GENETIC ASSOCIATION STUDIES AND THEIR CHALLENGES .................................................................. 13
1.2.1 COMMON DISEASE – COMMON VARIANT ............................................................................................ 13
1.2.2 REPLICATION .......................................................................................................................................... 14
1.2.3 MULTIPLE TESTING ............................................................................................................................... 15

1.3 PROSTATE CANCER ...................................................................................................................................... 15
1.3.1 INCIDENCE AND MORTALITY .................................................................................................................. 15
1.3.2 RISK FACTORS ........................................................................................................................................ 16
1.3.3 PROGRESSION .......................................................................................................................................... 18

1.4 GENETIC EPIDEMIOLOGY OF PROSTATE CANCER ................................................................................. 19
1.4.1 EPIDEMIOLOGICAL AND TWIN STUDIES ........................................................................................... 19
1.4.2 SEGREGATION ANALYSES ....................................................................................................................... 20
1.4.3 LINKAGE ANALYSES ............................................................................................................................... 20
1.4.4 CASE-CONTROL STUDIES ....................................................................................................................... 21
1.4.5 GENOME-WIDE ASSOCIATION STUDIES ............................................................................................. 21
1.4.6 THE FIRST CONFIRMED PROSTATE CANCER SUSCEPTIBILITY LOCUS .............................................. 22

1.5 CANDIDATE GENES IN PROSTATE CANCER ......................................................................................... 23
1.5.1 E-CADHERIN ......................................................................................................................................... 24
1.5.2 HORMONE REGULATING GENES .......................................................................................................... 24
1.5.3 ERG ......................................................................................................................................................... 25

2. AIMS ................................................................................................................................................................. 27

3. MATERIALS AND METHODS ...................................................................................................................... 28

3.1 DATA MATERIAL ......................................................................................................................................... 28
3.1.1 CAPS (PAPER I - V) .............................................................................................................................. 28
3.1.1.1 FOLLOW-UP (PAPER IV AND V) ........................................................................................................ 31
3.1.2 PROSTATE CANCER FAMILIES (PAPER I) ......................................................................................... 33

3.2 GENOTYPING METHODS .......................................................................................................................... 34
3.2.1 DASH (PAPER I) .................................................................................................................................. 34
3.2.2 SEQUENOM (PAPER II-V) ..................................................................................................................... 34
3.2.3 QUALITY CONTROL ............................................................................................................................... 35

3.3 SNP SELECTION ........................................................................................................................................... 35
3.3.1 E-CADHERIN (PAPER I) ....................................................................................................................... 35
3.3.2 REPLICATION STUDY (PAPER II) ......................................................................................................... 36
3.3.3 ANDROGEN PATHWAY GENES (PAPER III AND IV) ........................................................................ 38
3.3.4 ERG (PAPER V) .................................................................................................................................. 40

3.4 STATISTICAL METHODS .............................................................................................................................. 41
3.4.1 HAPLOTYPE TAGGING METHODS (PAPER I,III-V) ................................................................. 41
3.4.2 HARDY WEINBERG EQUILIBRIUM (PAPER I-V) ................................................................ 41
3.4.3 ASSOCIATION ANALYSIS (PAPER I-III, V) ......................................................................... 42
3.4.3.1 POLYMORPHISM ANALYSIS (PAPER I-III, V) ................................................................. 42
3.4.3.2 HAPLOTYPE ANALYSIS (PAPER I,III,V) ...................................................................... 42
3.4.4 TRANSMISSION/DISEQUILIBRIUM TESTING (PAPER I) .................................................... 42
3.4.5 SURVIVAL ANALYSIS (PAPER IV AND V) ........................................................................ 43
3.4.5.1 SNP ANALYSIS (PAPER IV AND V) .................................................................................. 43
3.4.5.2 HAPLOTYPE ANALYSIS (PAPER IV AND V) ................................................................. 43
3.4.6 ADJUSTMENT FOR MULTIPLE TESTING THROUGH PERMUTATION (PAPER I-III, V) .. 43
3.4.7 POPULATION ATTRIBUTABLE RISK (PAPER III) ............................................................... 43

4. RESULTS AND COMMENTS ........................................................................................................ 45
4.1 PAPER I ..................................................................................................................................... 45
4.2 PAPER II .................................................................................................................................... 46
4.3 PAPER III .................................................................................................................................. 49
4.4 PAPER IV ................................................................................................................................... 52
4.5 PAPER V .................................................................................................................................... 53

5. DISCUSSION .................................................................................................................................. 55
5.1 EVIDENCE OF GENETIC PREDISPOSITION TO PROSTATE CANCER (PAPER I-III, V) ....... 55
5.2 EVIDENCE OF GENETIC CONTRIBUTION TO PROSTATE CANCER PROGRESSION (PAPER IV AND V) .................................................................................................................. 56
5.3 GENETIC EPIDEMIOLOGY AND ASSOCIATION STUDIES – DESIGN, STRENGTHS AND LIMITATIONS .......................................................................................................................... 57
5.4 CAPS – DESIGN, STRENGTHS AND LIMITATIONS ............................................................... 59
5.5 STUDY DESIGN AND EXECUTION - MOLECULAR AND STATISTICAL METHODS .................. 63

6. SUMMARY AND CONCLUSIONS ................................................................................................. 65
6.1 FUTURE PROSPECTS – GENETIC EPIDEMIOLOGY OF COMPLEX DISEASES .................. 65
6.2 FUTURE PROSPECTS – GENETIC EPIDEMIOLOGY OF PROSTATE CANCER .................... 70
6.3 FUTURE IMPLICATIONS BASED ON THIS THESIS ................................................................ 71
6.4 CONCLUSIONS .......................................................................................................................... 72

7. POPULÄRVETENSKAPLIG SAMMANFATTNING (SUMMARY IN SWEDISH) ......................... 73

8. ACKNOWLEDGEMENTS ............................................................................................................... 75

9. REFERENCES .................................................................................................................................. 77
List of original publications


All publications are printed with permission from the publishers.
1. Background

Worldwide, more than 670,000 men are annually diagnosed with prostate cancer accounting for one ninth of all cancers in men (1). Prostate cancer incidence has increased dramatically since the introduction of PSA screening in the mid 90’s and it is currently the most common non-skin malignancy among men in industrialized countries (1). Prostate cancer is unique in its context with a high proportion of non-symptomatic latent cancers. Indeed, autopsy studies show that the absolute majority of elderly men do have small lesions of malignant cells in their prostate (2). On the other hand, a considerable fraction of prostate cancers has a rapid clinical course with a lethal disease outcome. In 2002, approximately 220,000 men died from prostate cancer worldwide making it the sixth most common cancer-related death in men (1). However, the risk for overtreatment is substantial considering the excellent prognosis of those with untreated localized prostate cancer (3) and the serious complications associated with radical treatment (4).

Despite considerable efforts little is known about prostate cancer aetiology and progression. Identifying the underlying mechanisms would be a major benefit for prevention, detection and treatment strategies. Most likely, prostate cancer is a consequence of the interplay between environmental and genetic factors. A large body of evidence suggests that inherited genetic variation is an important determinant of the probability to develop prostate cancer. This thesis aims to illuminate the possible role of a few candidate genes in prostate cancer aetiology and progression.

1.1 Genetic variation

1.1.1 Our human genome and its variability

The human genome consists of approximately 2.85 billion base pairs encoding roughly 20,000-25,000 genes (5). About 98% of our genomic DNA consists of non-coding regions comprising regulatory sequences and splicing sites but the vast majority is still obscure. Humans are essentially genetically identical but 0.1% of our genome differs between individuals (6). Genetic alterations are either inherited from parent to offspring and thereby present in each cell (germline mutations) or arise in a specific cell during cell division (somatic mutations). Genetic alterations may appear
as large-scale aberrations through translocations or loss and gains of chromosomes but the absolute majority takes place on a low-scale basis. The most frequent genetic alterations are single base substitutions but deletions and insertions of various lengths are also common. Alterations that take place within genes may affect the resulting protein in different ways: Nonsense mutations introduce a stop codon that truncates the protein sequence and missense mutations alter the amino acid sequence (7).

Recent discoveries suggest that copy number variants (CNVs) constitute a substantial fraction of the genetic diversity (8). CNVs include structural changes where a region of at least 1 kilobase (kb) of the genome is duplicated or deleted. Currently, more than 1,000 CNVs spanning approximately 143 megabases (Mb) of the genome have been detected in humans (8). It is clear that CNVs may alter gene expression and thereby contribute to phenotypic diversity. However, the importance of CNVs in evolution and disease mapping remains to be elucidated.

1.1.2 Single nucleotide polymorphisms

Single base substitutions that occur in more than 1% of the general population are called Single Nucleotide Polymorphisms (SNPs) (9). To date, 11.8 million SNPs corresponding to a SNP every 250th base pair in the human genome have been reported (10). Of those, about 7 million SNPs are common (i.e. have a minor allele frequency (MAF) >5%) (11). SNPs are by far the most abundant source of genetic variation in the human genome and they have several qualities that make them attractive in disease mapping (12). 1) They are distributed throughout the entire genome and reside in exons, introns, promoters, regulatory regions and between genes. 2) They are stable and easy to measure in the population. 3) Online SNP databases facilitate SNP identification and selection. 4) Advanced molecular techniques have made it economically and technologically feasible to perform genome-wide association studies assessing several hundred thousands of SNPs simultaneously.

1.1.3 Linkage disequilibrium

Linkage disequilibrium (LD) describes the non-random association between alleles for two or several genetic markers (13). LD arises when a new mutation occurs on a chromosome that carries a particular allele at a nearby locus. LD is influenced by several factors (14):
Recombination
Mutation
Genetic drift i.e. the random transmission of alleles from parent to offspring
Population admixture
Natural selection

There is great variation in LD across genomic regions but in general, LD decreases with increasing distance between markers (15). LD has become an important tool in genetic association studies as several markers can indirectly be measured through their allelic association with a genotyped variant (16).

1.1.4 Haplotypes and haplotype blocks

A haplotype is defined as a set of closely linked genetic markers along a chromosome which tend to be inherited together. Every individual carries one maternal and one paternal haplotype at each locus. Haplotypes have become an important tool in disease mapping as they complement single SNP analysis. Indeed, haplotype-based methods are suggested to be more powerful than single SNP analyses in indirect association studies (17).

In theory, $k$ SNPs in a region could generate $2^k$ distinct haplotypes but the number of observed haplotypes seldom exceeds $k+1$ (18). It appears as the genome can be divided into “blocks” of variable length where only a few common haplotypes are observed (Figure 1). These blocks are characterized by a pattern of strong LD and they are often punctuated by recombination hotspots (19). Several methods have been proposed for defining haplotype blocks. Broadly speaking: two main approaches exist: to retain LD or to limit haplotype diversity (20). Irrespective of method utilized the main determinants of block structure are SNP density, LD structure and source of population (21). For example, Africans tend to have higher haplotypic diversity than Caucasians (22).
**Figure 1:** The haplotype block concept. Blue and red represent two alternative alleles. 

(a) Haplotype blocks can be considered as short segments separated by recombination hotspots (zigzag lines). Within each block, there is little or no evidence for recombination and only a small number of distinct haplotypes is present. 

(b) Most chromosomes in the population are a mosaic arrangement of the variants within each block. Adapted from Cardon & Abecasis, Trends in Genetics, 2003 (23).

### 1.1.5 Haplotype tagging

The correlation between alleles in the genome makes it possible to capture common genetic variation within a region by genotyping only a limited number of “tagging” SNPs (tagSNPs). TagSNPs are selected either to capture the variation among other SNPs (tSNPs) or to capture the haplotypic variation (htSNPs).

In 2002, the International HapMap project launched the HapMap program (24). Their main goal is to construct a haplotype map of the human genome in order to define patterns of genetic variation across the human genome and to guide researchers in tagSNP selection. The HapMap project has so far genotyped 6.8 million SNPs distributed over the entire human genome except for chromosome Y. HapMap utilizes four different populations, 30 parent-offspring trios from the U.S. with European ancestry, 30 trios originating from the Yoruba population in Nigeria, 45 unrelated Japanese individuals from the Tokyo area and 45 unrelated individuals from Beijing, China. Complete genotype data from HapMap is publicly available online in order to facilitate haplotype tagging for individual researchers. It has been shown that HapMap data is sufficient to capture the common variation (i.e. >5%) in
the genome. However, original density of SNPs, minor allele frequency (MAF) and LD structure may have a substantial impact on achieved genetic coverage (25,26).

1.2 Genetic association studies and their challenges

To date, most genetically characterized human disorders are mendelian. Mendelian traits have predominantly been mapped using linkage analysis but for polygenic diseases this method has had limited success (16). Instead, association studies have been proposed as a more powerful alternative to identify susceptibility genes in complex diseases (16,27).

The aim of genetic association studies is to identify polymorphisms associated with a given trait on a population level. Along with the mapping of the human genome and the development of advanced genotyping technologies, association studies have become increasingly popular (28). Association can be directly tested for a single putative causal marker but if no strong priori for a specific marker exists, indirect LD-based association may be preferable. In this case, a number of polymorphisms acting as surrogates for the genetic variation are genotyped and tested for association (tagSNPs) (29). However, the usefulness of association studies in identifying susceptibility genes have been debated (30).

Recent advances in molecular techniques have made it possible to perform large-scale genotyping of several hundred thousands of SNPs simultaneously. Such genome-wide association studies have already resulted in the mapping of susceptibility genes for several complex diseases including diabetes (31-35), cancer (36-42) and coronary heart disease (43,44). Most likely, genome-wide association studies will result in an explosion of identified susceptibility loci for complex diseases in the near future.

1.2.1 Common disease – common variant

By nature, most monogenic disorders are infrequent and therefore have a limited significance on a population level. Consequently, the aetiology of common diseases with a heritable component is likely to involve multiple genetic variants and environmental factors. Various possibilities are consistent with a polygenic disease model, ranging from a few susceptibility genes with moderate risk effects to a large number of alleles each contributing with only a slight risk increase. Indeed,
experience from identified susceptibility variants in complex diseases suggests low-penetrant genes having odds ratios (ORs) in the order of 1.1-1.5 (45).

Since many complex diseases are common, they are most likely caused by genetic variants that are common in the population (46), also referred to as the Common Disease –Common Variant (CD-CV) hypothesis. One example of a common genetic variant consistently associated with a complex disease is the PPARγ gene in type 2 diabetes (47). Although this variant only confers a slight risk increase (OR 1.25), its high prevalence (MAF 15%) causes a relatively high attributable risk in the population. A proposed alternative to the CD-CV hypothesis is the genetic heterogeneity model which states that multiple rare variants at different loci are individually sufficient to cause disease (48). Due to their low frequency and the number of sufficient causing loci these variants would most likely not be detected in an association study (12).

1.2.2 Replication

Genetic association studies have been plagued by a high rate of false positive associations. More than 600 potential associations between common variants and disease had been reported in 2002 (49). Hirschhorn and colleagues reviewed 166 of these associations and observed that only six of them were consistently replicated. There are several possible reasons explaining the lack of replication in genetic association studies (28):

- False positive findings caused by random fluctuations
- False negative findings caused by lack of statistical power
- True variation in disease-causing alleles between different populations
- Population stratification
- Different LD patterns in different populations in case of indirect association
- Gene-gene and gene-environment interactions
- Difference in disease classification
- Genotyping errors

Minimizing spurious findings and false negative results demands qualitative study designs. A well characterized phenotype, a large homogenous population, validation of the genotyping procedure and appropriate statistical analyses are all crucial
components that must be considered in interpreting results from a genetic association study.

1.2.3 Multiple testing

Along with the increasing number of association studies and the possibility of analyzing many genetic variants at low cost, the problem of multiple testing has become more prominent. To address this issue several methods have been proposed. A common approach is to use Bonferroni correction where the crude p-value is multiplied with the number of performed tests. However, this method may not be appropriate in genetic association studies with tightly linked markers as Bonferroni correction does not consider the correlation between genetic variants (50).

A more accurate approach is to use a permutation procedure which considers the observed LD structure in the region of interest (51). New data sets are generated by shuffling phenotypic status between all individuals while genotypes are retained under the null hypothesis of no association. Association between genotypes and phenotype is then assessed and the maximum test statistic is drawn for each permuted data set. Adjusted p-values are then obtained from the empirical distribution of all maximum test statistics.

It has been argued that the prior probability rather than number of SNPs tested should be considered when validating an association. If no or few prior beliefs exist, the statistical evidence of association has to be strong in order to be considered as true. This approach is known as the Bayesian paradigm (52). Wacholder and colleagues combine Bayesian ideas with classical statistics and propose assessment of the false positive report probability (FPRP) which considers both the prior probability and statistical power when evaluating an association (53).

1.3 Prostate cancer

1.3.1 Incidence and mortality

Prostate cancer is the most common malignancy among men in Sweden, accounting for one third of all male cancer cases. In 2005, 9,881 Swedish men were diagnosed with prostate cancer and 2,549 prostate cancer-specific deaths were reported, making it the leading cause of cancer death among men in Sweden (54). Although prostate cancer incidence has increased rapidly during the last decades prostate cancer
mortality has remained constant (Figure 2). The increase in incidence rate is largely due to the introduction of PSA testing among asymptomatic individuals resulting in a higher proportion of cancers that are still confined in the prostate.

![Graph of incidence and mortality rates of prostate cancer in Sweden 1958-2004.](image)

**Figure 2**: Incidence (solid) and mortality (dotted) rates of prostate cancer in Sweden 1958-2004.

About 679,000 new cases of prostate cancer and 220,000 prostate cancer deaths were reported worldwide in 2002. There is a wide variation in incidence rates between countries. United States, Northern and Western Europe, Australia and New Zealand have the largest incidence rates whereas Asia has the lowest (1).

### 1.3.2 Risk factors

The single most important risk factor for prostate cancer is age. Fifty percent of all incident cases in Sweden 2005 were older than 70 years at time for diagnosis and the mean age of diagnosis between 1996 and 2000 was 74 years. Incident cases younger than 45 years are rare (54).

**Ethnicity**

There are large differences in prostate cancer incidence between ethnicities worldwide. The highest incidence is found in African-Americans with a 60 times higher rate compared to Shanghai, China, where rates are lowest (55). Studies have
shown that migrants from low-risk to high-risk countries experience an increase in incidence suggesting that environmental and lifestyle factors are important in prostate cancer aetiology. However, there is a marked difference in incidence rates among different ethnicities living in the same region. In United States, men with Asian descent have the lowest rates, Caucasians intermediate and African-Americans the highest. These findings suggest that differences between ethnic populations are genuine and not explained only by differences in health care and life-style factors (56).

Family history
Familial aggregation of prostate cancer has repeatedly been observed and accumulating evidence exists that heredity is an important risk factor. Approximately 10–15% of men with prostate cancer have reported at least one relative who is also affected (57). Hereditary prostate cancer is further discussed in section 1.4.

Diet
The possible involvement of dietary factors in prostate cancer aetiology is controversial. Dietary intake of processed tomatoes has consistently been associated with a lower risk of developing prostate cancer, mainly attributed to lycopene (58). Selenium has been associated with a reduced risk of prostate cancer in randomized controlled clinical trials as well as prospective and retrospective studies (58). Fatty fish intake has consistently been associated with a reduced risk of prostate cancer both in prospective and retrospective studies (59-61). Moreover, phytoestrogens or soy products have been shown to reduce prostate cancer risk in a Swedish population (62) but worldwide studies have been inconsistent in their findings (63). Fat, red meat, zinc, calcium, vitamin D and tea have all been suggested to alter prostate cancer risk but without any overall support (63).

Inflammation
There is emerging evidence that inflammation plays an important role in prostate cancer (64). Aspirin use has repeatedly been associated with reduced prostate cancer risk and a recent meta-analysis identified a 10% risk reduction for aspirin users (95% CI: 0.8-1.0) (65). The presence of prostatitis has been associated with prostate cancer risk (pooled OR 1.65, 95% CI: 1.0-2.5) (66) but these studies are based on self-reported
assessment of prostatitis and should therefore be interpreted with caution. In addition, a history of sexually transmitted infection has been associated with prostate cancer risk in a meta-analysis (OR 1.4, 95% CI: 1.2-1.7) (67) but the included studies had limited study samples and were therefore difficult to interpret (64). Genetic variability in inflammatory pathways has been implicated to influence prostate cancer susceptibility (68) and inflammatory genes including COX-2, IL1RN, MIC-1, TLR-4 and the TLR-1-6-10 gene cluster have all been associated with prostate cancer risk in a Swedish population (69-72).

**Hormones**

It has been more than 60 years since Huggins and Hodges demonstrated that androgen ablation causes tumor regression for metastatic prostate cancer (73). Huggins received the Nobel Prize in 1966 for his findings and today, androgen therapy is the standard treatment for advanced prostate cancer. The growth and maintenance of the prostate depend on androgens as they stimulate proliferation and inhibit apoptosis. Despite the impact of androgens in prostate cancer, data from 16 prospective studies addressing the association of circulating levels of androgens and prostate cancer risk have not provided any evidence of a relationship (74). In addition, there is no clear relationship between circulating levels of androgens and the androgenic action in the prostate gland.

**Other risk factors**

Several other risk factors including obesity, physical activity, smoking and alcohol use have been implicated in prostate cancer but there is overall no support for their involvement in prostate cancer aetiology (58).

**1.3.3 Progression**

The relative five-year survival for prostate cancer in Sweden has increased from 37% to 78% during the last three decades (54). Still, it is the leading cause of cancer-related death among men in Sweden. One of the toughest challenges in prostate cancer management is to identify those men who are in need of an aggressive treatment. A substantial fraction of diagnosed men have a clinically insignificant cancer that will never progress to become life-threatening. In fact, autopsy studies have shown that 30% of men in their 30s, 50% of men in their 50s, and more than 75% of men older than 85 years have a latent prostate cancer (75). Traditional prognostic factors such as
PSA serum levels, TNM stage and Gleason sum explain only a limited fraction of the variability in outcome. This emphasizes the importance of detecting new independent prognostic markers in order to improve treatment strategies.

Tumor development and progression are a consequence of somatic genetic alterations that activates proto-oncogenes and inhibits tumor-suppressor genes. These mechanisms induce uncontrolled cellular proliferation, aberrant apoptosis and consequently invasion of malignant cells to other tissues. Even though somatic mutations are the key factors of cancer progression it has been suggested that germline genetic variation has a considerable impact on progression (76). Due to its strong influence from genetic components on its incidence, prostate cancer has been suggested as an appealing target for identification of inherited factors that contribute to progression (77). A recent study from Iceland investigated the 999del5 founder mutation in BRCA2 in 527 prostate cancer cases and concluded that carriers had a significant risk excess of dying from prostate cancer (Hazard ratio (HR): 3.42, 95% CI: 2.1-5.5) (78). There are a few other studies that have investigated germline genetic variation and prostate cancer-specific death but only in limited study materials (79-82).

1.4 Genetic epidemiology of prostate cancer

There is a large body of evidence supporting the involvement of genetics in prostate cancer aetiology. Various study designs including twin studies, family-based studies, and case-control studies all suggest the importance of genetics but to date there has been little progress in identifying the responsible genes. The breast cancer susceptibility gene BRCA2 seems to alter prostate cancer risk but it contributes to less than 5% of the hereditary and young onset cases (age <55) (83). A major step forward was taken last year as a susceptibility locus on chromosome 8q24 was identified in a linkage analysis from Iceland (84). This locus has now been replicated in numerous populations and is considered the first convincing low-penetrant prostate cancer locus (see section 1.4.6).

1.4.1 Epidemiological and twin studies

Case-control studies and cohort studies have consistently reported a two-three fold risk increase for men who have at least one first degree relative diagnosed with prostate cancer. A recent meta-analysis on 13 case-control and cohort studies
estimated a pooled odds ratio of 2.5 (95% CI: 2.2-2.8) for men with one first-degree relative and 3.5 (95% CI: 2.6-4.8) for men with two affected first-degree relatives (85). In general, studies have reported a correlation between prostate cancer risk and number of relatives affected. General observations about prostate cancer and family history include higher risk for men with an affected brother compared to those with an affected father and declining importance of heredity with age. A twin study originating from Scandinavia estimated the hereditable impact on 28 different tumor types and concluded that prostate cancer has the strongest hereditability (42%, 95% CI: 29-50) among all common malignancies (86).

1.4.2 Segregation analyses

The aim of segregation analyses is to establish the mode of inheritance of a specific trait within families. To date, nine segregation analyses on prostate cancer have been published. The majority of studies propose an autosomal dominant inheritance model with a high-penetrant rare allele. However, autosomal recessive, X-linked and multifactorial inheritance patterns have also been suggested. The proposed autosomal dominant model should be considered with caution as the collection of families are often based on the Carter Criteria (87) which itself is biased towards an autosomal dominant inheritance. It is difficult to compare these studies as they differ in country of origin, proband ascertainment and statistical methods.

1.4.3 Linkage analyses

Multiple prostate cancer linkage studies have been conducted but findings have been inconclusive and irreproducible. The first study to report significant linkage was undertaken in the United States 1996 where a region at chromosome 1q24-25 (HPC1) was identified (88). Since then, numerous loci have been suggested and refuted, demonstrating the complexity of prostate cancer genetics. Linkage analyses in prostate cancer suffer from recruitment difficulties due to late age of onset, phenocopies due to sporadic cases within families and locus heterogeneity. In an attempt to overcome these problems research groups around the world joined forces and formed “The International Consortium for Prostate Cancer Genetics” (ICPCG) with the aim to identify prostate cancer susceptibility loci through combined analyses of linkage data. In a genome-wide linkage analysis using 1,233 families they identified 5 suggestive but not statistically significant loci: 5q12, 8p21, 15q11, 17q21 and 22q12 (89). Amundadottir and colleagues performed a linkage analysis using
1,068 microsatellite markers among 323 prostate cancer families and identified suggestive linkage (LOD score 2.11) at chromosome 8q24 (84). Subsequent studies have confirmed this finding and it is now recognized as the first low-penetrant prostate cancer susceptibility locus (section 1.4.6).

1.4.4 Case-control studies
Hundreds of genetic association studies on prostate cancer have been conducted during the last decade. The vast majority of studies have investigated single variants in genes that are believed to be involved in important pathways for malignant transformation of prostate cells such as hormone and cell cycle regulation or inflammation mechanisms. The introduction of the haplotype tagging concept has led to more comprehensive studies of the genetic variation within a region of interest but most often studies lack sufficient statistical power due to small sample sizes. To overcome this issue, several cohorts joined forces in 2003 and formed “The National Cancer Institute Breast and Prostate Cancer Cohort Consortium” (BPC3) which include data from nine prospective cohorts. In total, BPC3 comprises approximately 9,000 prostate cancer cases and 10,000 controls. BPC3 aims to characterize genetic variation in approximately 50 genes that mediate steroid-hormone metabolism and insulin-like growth factor signalling (90). So far, the consortium has published two papers regarding prostate cancer and genetic variation. The first paper investigated the HSD17B1 gene and found no overall association (91). The other study replicated the rs1447295 SNP located on chromosome 8q24 previously identified by linkage analysis (92).

1.4.5 Genome-wide association studies
Recently, two independent genome-wide SNP scans in prostate cancer were published. The Cancer Genetic Markers of Susceptibility (CGEMS) project performed a scan of 550,000 SNPs in 1,172 cases and 1,157 controls of European origin (41). They identified a novel region of 8q24 located 70 kb away from the rs1447295 SNP identified by the Icelandic linkage study. Combined analysis confirmed association for the rs6983267 SNP in four additional populations comprising a total of 4,296 cases and 4,299 controls (OR=1.26, 95% CI: 1.1-1.4 for heterozygote carriers, P=9.4·10⁻¹³).

At the same time, Gudmundsson and colleagues (42) genotyped 317,000 SNPs in 1,453 prostate cancer cases and 3,064 controls originated from Iceland. They observed the strongest association in the 8q24 region. No other SNP reached
genome-wide statistical significance. Subsequent haplotype analysis revealed a second region at chromosome 8q24 located 300 kb upstream from the original rs1447295 SNP. One single haplotype (HapC) was associated with a two-fold risk (95% CI: 1.8-2-5, \(P=3.1 \times 10^{-15}\)) of developing prostate cancer in a combined analysis from four independent populations with European ancestry.

Gudmundsson and colleagues followed up their initial genome-wide scan by increasing the number of cases to 1,501 and controls to 11,290 (37). No new locus reached genome-wide significance but based on prior linkage findings on chromosome 17q they explored this region further as six SNPs in this region were among the 100 most significant in their genome-wide scan. They identified three SNPs that were associated with prostate cancer in four independent populations with European ancestry comprising in total 3,490 cases and 14,345 controls. All three SNPs showed moderate risk effects with odds ratios of \(\sim 1.2\) in all populations combined.

Interestingly, one of the SNPs is located in the TCF2 gene and showed a protective effect against type 2 diabetes in eight case-control studies of European, African and Asian ancestry. This finding may in part explain the inverse relationship between prostate cancer risk and type 2 diabetes that has been observed (93).

1.4.6 The first confirmed prostate cancer susceptibility locus

In 2006, an Icelandic group identified suggestive linkage (LOD score 2.1) at chromosome 8q24 among 323 prostate cancer families (84). They refined the region by genotyping additional markers in a case-control study consisting of 869 sporadic prostate cancer cases and 596 population controls. The strongest association was observed for the DG8S737 microsatellite where carriers of the -8 allele had an almost two-fold risk to develop prostate cancer. They also observed association for a nearby located common SNP, rs1447295 (MAF=11%) in two populations with European ancestry originating from Sweden (OR: 1.29, \(P=0.0045\)) and United States (OR: 1.66, \(P=0.0067\)). They also observed association for the DG8S737 marker in an African-American population comprising 246 cases and 352 controls (OR: 1.60, \(P=0.002\)) but not for the rs1447295 SNP (OR: 1.15, \(P=0.29\)). Combining results from the different populations revealed an OR of 1.62 \((P=3 \times 10^{-11})\) for DG8S737 -8 and an OR of 1.51 \((P=1 \times 10^{-11})\) for rs1447295. Simultaneously with this work, an American group performed a genome-wide admixture scan in 1,597 African Americans and identified a 3.8 Mb region at chromosome 8q24 (94). They replicated the Icelandic findings but observed that the previously described alleles only explained a fraction of their
admixtue signal. They followed up their findings by genotyping additional 2,973 SNPs in up to 4,266 prostate cancer cases and 3,252 controls, and identified five novel genetic variants spanning a region of 430 kb of 8q24 associated with prostate cancer risk (95).

To date, seven studies have confirmed the association between prostate cancer risk and the 8q24 variant rs1447295 (Table 1). At least three risk-associated regions on 8q24, separated by sites of recombination, confer independent risk. Recent studies demonstrate that genetic variation at 8q24 is also involved in development of colon cancer (39, 40, 96, 97) and breast cancer (36) suggesting that 8q24 is important for cancer development in general. 8q24 is a gene-poor region with high recombination rate. The proto-oncogene c-MYC lies approximately 260 kb telomeric of rs1447295 but subsequent analyses have shown no evidence of association between c-MYC variation and prostate cancer risk (41,42). Thus, the biological mechanisms depending on genetic variation at 8q24 are still elusive.

Table 1: Association between prostate cancer risk and the rs1447295 SNP located at 8q24

<table>
<thead>
<tr>
<th>Study</th>
<th>Population</th>
<th>Number of Cases/Controls</th>
<th>MAF (%)</th>
<th>Odds Ratio (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amundadottir (84)</td>
<td>Icelandic</td>
<td>1,291/997</td>
<td>11</td>
<td>1.72</td>
<td>1.7·10⁻⁹</td>
</tr>
<tr>
<td>Swedish</td>
<td>1,435/779</td>
<td>13</td>
<td>1.19</td>
<td>0.0045</td>
<td></td>
</tr>
<tr>
<td>US Caucasians</td>
<td>458/247</td>
<td>8</td>
<td>1.66</td>
<td>0.0067</td>
<td></td>
</tr>
<tr>
<td>Afro-Americans</td>
<td>246/352</td>
<td>31</td>
<td>1.15</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Freedman (94)</td>
<td>Multiethnic cohort</td>
<td>1,614/1,547</td>
<td>10-17</td>
<td>1.36 (1.2–1.5)</td>
<td>4.2·10⁻⁹</td>
</tr>
<tr>
<td>Severi (98)</td>
<td>Australian</td>
<td>821/731</td>
<td>11</td>
<td>1.52 (1.2-1.9)</td>
<td>0.0005</td>
</tr>
<tr>
<td>Suuriniemi (99)</td>
<td>US Caucasians</td>
<td>597/548</td>
<td>11</td>
<td>1.43 (1.1-2.0)</td>
<td></td>
</tr>
<tr>
<td>Schumacher (92)</td>
<td>US Caucasians</td>
<td>5,961/6,718</td>
<td>11</td>
<td>1.34 (1.2-1.5)²</td>
<td>1.2·10⁻¹³</td>
</tr>
<tr>
<td>Wang (100)</td>
<td>US Caucasians</td>
<td>435/545</td>
<td>10</td>
<td>1.16 (0.9–1.6)</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>US Caucasians</td>
<td>491/545</td>
<td>10</td>
<td>1.93 (1.4–2.7)</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>US Caucasians</td>
<td>195/545</td>
<td>10</td>
<td>1.87 (1.3–2.7)</td>
<td>0.0005</td>
</tr>
<tr>
<td>Gudmundsson (42)</td>
<td>Icelandic</td>
<td>1,453/3,064</td>
<td>10</td>
<td>1.71 (1.5-2.0)</td>
<td>1.6·10⁻¹⁴</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>385/892</td>
<td>7</td>
<td>1.44 (1.1-1.9)</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>367/1,302</td>
<td>11</td>
<td>1.30 (1.1-1.8)</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>US Caucasians</td>
<td>458/251</td>
<td>8</td>
<td>1.56 (1.1-2.3)</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>African Americans</td>
<td>373/372</td>
<td>31</td>
<td>1.01 (0.8-1.3)</td>
<td>0.96</td>
</tr>
</tbody>
</table>

¹Depending on ethnicity, ²Heterozygous carriers

1.5 Candidate genes in prostate cancer

Numerous genes have been implicated to alter prostate cancer susceptibility. The most likely candidates are believed to be involved in the metabolism of sex-steroid hormones (101). Although several candidate genes have been associated with prostate cancer risk it has proven difficult to replicate earlier findings in a consistent manner. In this thesis, we have investigated the tumor suppressor gene E-cadherin (CDH1), the androgen regulating genes androgen receptor (AR), cytochrome P450...
(CYP17) and steroid-5-alpha-reductase (SRD5A2) and the proto-oncogene v-ets erythroblastosis virus E26 oncogene homolog (ERG).

1.5.1 E-cadherin

E-cadherin (CDH1) is a cell-cell adhesion molecule expressed in epithelial tissues. It is recognized as an invasion-suppressor gene as aberrant gene expression may cause dysfunctional cell-cell adhesion and thereby induce local invasion. CDH1 maps to 16q which is characterized by loss of heterozygosity in prostate cancer (102-106). Epigenetic events have been observed in CDH1 in prostate tumors and the degree of methylation has been correlated with disease severity (107). Furthermore, aberrant CDH1 expression has been associated with malignant transformation of the prostate and metastasis (108). A single base substitution located in the promoter region of CDH1, rs16260, has shown to affect the transcription activity with a 68% reduction for carriers of the variant allele (109). This SNP has been associated with prostate cancer risk in several studies but others have not observed any association (110-115).

1.5.2 Hormone regulating genes

Due to the importance of androgens in prostate cancer, genes involved in the sex-steroid hormones metabolism have been proposed as plausible candidate genes in prostate cancer. The androgen receptor (AR) is a transcription factor which binds to androgen responsive elements located in target genes and thereby activates their transcription. Metastatic prostate cancer is primarily treated with either androgen blockade to reduce levels of circulating androgens or with AR antagonists that reduce AR activity (116). AR is by far the most investigated gene in prostate cancer. The majority of studies have investigated a poly-glutamine (CAG) repeat located in the transactivation domain. The length of this microsatellite varies typically between 18 and 22 repeats and few repeats have been associated with a higher transcriptional activity (117,118). A recent meta-analysis based on 19 studies concluded that the odds ratio per repeat decrement was 1.02 (95% CI: 0.99-1.06) (119). A polymorphic GGN repeat located downstream of the CAG repeat has also been proposed as a prostate cancer susceptibility polymorphism. However, combined results from eight different studies yielded an odds ratio of 1.01 (95% CI: 0.98-1.04) per repeat decrement (119). Remarkably few studies have assessed other polymorphisms in AR. Freedman and co-workers conducted a comprehensive evaluation of the gene by testing 32 SNPs for association in a study comprising 1,756 individuals. Although
they did not find any overall significant results, 11 SNPs were associated with advanced prostate cancer (120).

AR variation has also been implicated as a determinant of prognosis for prostate cancer patients. Early prostate cancer is androgen-dependent and androgen withdrawal results initially in tumor regression but eventually hormone-refractory cancer cells emerge signalling that prognosis has become poor (121). However, the duration of response to hormonal therapy varies widely. Although the causes for this variation remain unknown, several experimental observations suggest that AR is involved (122). AR expression increases during hormonal treatment failure and AR amplification is found in one third of hormone-refractory tumor cells (123). A few studies have assessed the relationship between prognosis and number of CAG repeats with PSA recurrence, failure of hormonal response or disease-specific survival as primary end-points with inconsistent results (124-130).

The SRD5A2 gene encodes for the 5α-reductase type 2 enzyme which catalyses the conversion of testosterone to the more potent androgen dihydrot testosterone. It has been shown that Japanese men have lower 5α-reductase activity than US Caucasians and Africans suggesting a possible explanation of the difference in prostate cancer incidence between ethnicities (131,132). Two SNPs located at codon 49 (rs9282858) and codon 89 (rs523349) have been shown to alter the enzyme activity (133). These polymorphisms have been extensively investigated for association with prostate cancer risk but with equivocal findings (134).

CYP17 encodes for the P450c17α enzyme which catalyzes two key mechanisms in the steroid biosynthesis pathway. A single SNP in the promoter region of CYP17, rs743572, has been proposed to increase the synthesis of androgens but subsequent studies have not been able to confirm this observation (135-137). A meta-analysis based on 12 case-control studies including altogether >5,000 individuals examined the relationship between rs743572 and prostate cancer risk and found no evidence of association (OR=1.08, P=0.24) (138).

1.5.3 ERG

An important event in cancer development is chromosomal translocations where either two genes are merged resulting in a new protein or where an enhancer/promoter sequence is juxtaposed to a proto-oncogene activating the latter. These chromosomal rearrangements are common in lymphomas (e.g. t(8;14) resulting in activation of the c-MYC oncogene) and leukaemias (e.g. t(9;22) also
known as the Philadelphia chromosome) (139) but are rare events in solid tumors. The importance of identifying these genetic aberrations can be illustrated by the Philadelphia chromosome where a molecular targeted drug (Gleevec®) has been developed. Gleevec® inhibits the activity of the new protein created by the fusion and is now offered to cancer patients who carry the Philadelphia chromosome.

Recently, Tomlins and colleagues scanned common cancers for oncogenes with aberrant expression in order to identify chromosomal translocations (140). Two proto-oncogenes, ERG and ETV1 were identified as expression outliers in prostate cancer and further analysis showed that the promoter region of the androgen-dependent gene TMPRSS2 (21q22.3) merged with ERG (21q22.3) or ETV1 (7p21.2). Subsequent studies from several groups has confirmed that the TMPRSS2:ERG fusion is prevalent in up to 50% of tumors (141-147). Genomic translocations involving ERG in cancer is not a novel event as it has been observed in Ewing’s sarcoma and several types of myeloid leukaemia (148).

Presence of the ERG:TMPRSS2 fusion has been associated with increased prostate cancer-specific death in a watchful waiting cohort (HR=2.7; 95% CI: 1.3-5.8) (149) but this was not confirmed in another study (143). Interestingly, heterogeneity in the ERG:TMPRSS2 fusion due to alternative fusion sites in the genome has been observed. Depending on subtype, different clinical characteristics such as Gleason sum and tumor size have been observed, suggesting that the presence and nature of this fusion is indeed important for prognosis (146,150). If germline genetic variation at the target region for the fusion affects the ability to fuse or the specific fusion site, it could alter prostate cancer risk and foremost progression.
2. Aims

The general aim underlying the work in this thesis is to elucidate the role of inherited genetic variation in prostate cancer aetiology and progression.

The specific aims of this thesis are:

- To explore if common genetic variation in E-cadherin alters prostate cancer risk, especially for men with a positive family history of prostate cancer.

- To replicate earlier reported genetic associations in prostate cancer.

- To investigate if common genetic variation in three key regulating genes in the androgen synthesis (AR, SRD5A2 and CYP17) affects prostate cancer risk, progression and hormonal therapy response.

- To examine if common genetic variation at the ERG locus affects prostate cancer risk and outcome.
3. Materials and methods

3.1 Data material

3.1.1 CAPS (Paper I-V)

All studies in this thesis are based on CAPS (CAncer Prostate in Sweden), a population-based prostate cancer case-control study. The study base for CAPS included all men between 35 and 79 years living in the central and northern part of Sweden and all men between 35 and 65 years living in the south-eastern part of Sweden and Stockholm. Approximately 6 million people (67% of the Swedish population) reside within this recruitment area (Figure 3).

![Figure 3: Recruitment area for the CAPS study.](image)

We identified all incident prostate cancer cases between March 2001 and October 2003 through four out of six existing regional cancer registries in Sweden. As Swedish law requires that both the attending physician and pathologist report newly diagnosed cancer cases to the cancer registries, they comprise almost 100% of all
diagnosed cancers. An administrator at the regional cancer registry mailed a letter to the attending physician to inform her/him about the CAPS study and asked for permission to invite the patient. If the physician approved, she/he mailed a letter to the patient describing the study and asked the patient to participate by returning a mail. Data collection was accomplished in two separate rounds (CAPS1 and CAPS 2). For patients diagnosed between March 2001 and September 2002 (CAPS1) a comprehensive self-administrated questionnaire regarding diet, family history of prostate cancer, physical activity and smoking were used whilst a more concise questionnaire with questions regarding family history of prostate cancer, prostatitis, non-steroid anti inflammatory drugs and aspirin use were used for patients diagnosed between October 2002 and October 2003 (CAPS2).

A total of 3,648 men were invited and of them 3,161 (87%) agreed to participate by answering the questionnaire and/or donate a blood sample. We obtained blood from 2,965 (81%) of the invited cases. Through linkage to the National Prostate Cancer Register we obtained clinical information about TNM-stage, differential grade, Gleason sum, PSA serum levels and primary treatment at diagnosis (Table 2). A case who met one of the following conditions was classified as an advanced prostate cancer case prone to progression (1,264 patients): T3/4, N+, M+, GIII, Gleason sum 8–10 and PSA serum level ≥50 μg/L. If none of these conditions were fulfilled the patient was classified as a localized prostate cancer (1,701 cases). The mean age for case patients was 66.3 years (range: 45-82).

If a case subject in CAPS reported that he had at least one relative diagnosed with prostate cancer he was followed up by a second self-administrated questionnaire and a subsequent telephone interview by a research nurse. All reported prostate cancer diagnoses in first-, second-, and, if possible, third-degree relatives were verified through the Cancer Register or medical records. A patient was classified as a hereditary prostate cancer (HPC) if he fulfilled the Carter Criteria of a hereditary prostate cancer (87) which requires three or more relatives affected with prostate cancer in any nuclear family, prostate cancer in three successive generations in either of the probands’ paternal or maternal lineages or two first-degree relatives affected with prostate cancer at 55 years of age or younger. If a patient had one first-degree relative affected with prostate cancer he was classified as a familial prostate cancer (FPC). In total, CAPS consists of 2,862 cases with no family history, 206 FPC cases and 87 HPC cases.
Table 2: Clinical characteristics of case subjects who participated with a blood sample in the CAPS study.

<table>
<thead>
<tr>
<th>CHARACTERISTICS</th>
<th>N= 2,965</th>
<th>CHARACTERISTICS</th>
<th>N= 2,965</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status at follow-up¹</td>
<td></td>
<td>T stage</td>
<td></td>
</tr>
<tr>
<td>Alive</td>
<td>2,429</td>
<td>T0/TX</td>
<td>83</td>
</tr>
<tr>
<td>Deceased from other events</td>
<td>174</td>
<td>T1</td>
<td>1,106</td>
</tr>
<tr>
<td>Deceased from prostate cancer</td>
<td>362</td>
<td>T2</td>
<td>928</td>
</tr>
<tr>
<td>Differential Grade</td>
<td></td>
<td>T3</td>
<td>737</td>
</tr>
<tr>
<td>GI</td>
<td>160</td>
<td>T4</td>
<td>111</td>
</tr>
<tr>
<td>GII</td>
<td>601</td>
<td>N stage</td>
<td></td>
</tr>
<tr>
<td>GIll</td>
<td>327</td>
<td>N0/NX</td>
<td>2,867</td>
</tr>
<tr>
<td>Missing</td>
<td>1,877</td>
<td>N1-N3</td>
<td>98</td>
</tr>
<tr>
<td>Gleason Sum</td>
<td></td>
<td>M stage</td>
<td></td>
</tr>
<tr>
<td>≤ 4</td>
<td>108</td>
<td>M0/MX</td>
<td>2,677</td>
</tr>
<tr>
<td>5</td>
<td>298</td>
<td>M1</td>
<td>288</td>
</tr>
<tr>
<td>6</td>
<td>1,007</td>
<td>PSA levels, ng/ml</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>811</td>
<td>&lt; 4</td>
<td>153</td>
</tr>
<tr>
<td>8</td>
<td>262</td>
<td>4-9.99</td>
<td>1,025</td>
</tr>
<tr>
<td>9</td>
<td>194</td>
<td>10-19.99</td>
<td>671</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>20-49.99</td>
<td>475</td>
</tr>
<tr>
<td>Missing</td>
<td>260</td>
<td>≥ 100</td>
<td>235</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Missing</td>
<td>326</td>
</tr>
</tbody>
</table>

¹Date of last follow-up: March 1st, 2007.

We selected control subjects randomly from the Swedish Population Register. Controls were frequency matched to cases according to the expected age distribution (groups of five-year intervals) and geographic origin (two areas representing northern and southern part of Sweden). All selected controls received an introduction letter describing the study and three to four weeks later they received the same self-administered questionnaire used for the cases. A total of 3,153 controls subjects were invited, 2,149 (68%) agreed to participate and of them, 1,823 (58%) donated a blood sample. Both cases and controls were re-contacted three times to improve response rate. The mean age for participating controls was 67.2 years (range: 45-80).

All CAPS participants were asked to donate four 10 ml blood samples at their closest health centre or hospital. The blood samples were then sent by overnight mail to the Medical Biobank in Umeå. Upon arrival, they were separated into serum, plasma, leukocytes and erythrocytes. Samples were then stored at -70°C until analysis. Genomic DNA was extracted from leukocytes using standard techniques.
3.1.1.1 Follow-up (Paper IV and V)

In order to assess risk factors for prostate cancer-specific death we retrieved information on date and cause of death for all case subjects in CAPS through linkage to the Swedish Causes of Death Register (54). Each study participant is identified through his individually unique national registration number which includes date of birth. Using this registration number, follow-up for prostate cancer-specific mortality was achieved up until July 15th, 2006 for paper IV and March 1st, 2007 for paper V. We defined prostate cancer-specific death as those who had prostate cancer classified as the underlying cause of death. If a person was not found in the register he was assumed to be alive. For those who died after December 31st, 2004 we acquired cause of death certificates and let an experienced oncologist review them. Based on updated data in March, 2007 our own classification of cause of death could be evaluated for those who deceased during 2004. A total of 138 cases deceased in 2004 and of them, seven were classified differently in the register leading to a 95 % agreement between our own classification and the register. The average time of follow-up was 4.4 years (0.3-6.5). In total, 576 (18%) prostate cancer cases in CAPS were deceased at March 1st, 2007 and of them 398 (13%) had prostate cancer as the underlying cause of death. As expected, Gleason sum, PSA serum levels at diagnosis, tumor grade and presence of lymph node or metastases all strongly correlate with survival ($P<0.00001$, Figure 4).
Figure 4: Prostate cancer-specific survival curves for known prognostic factors in prostate cancer based on 3,161 prostate cancer cases included in the CAPS study. a) Gleason sum, b) PSA at diagnosis, c) T-stage, d) NM-stage.
To validate our categorization into advanced and localized cases (where a subject is classified as an advanced case if he met at least one of the following criteria: T3/T4, N+, M+, Gleason score of 8-10 or PSA level ≥ 50 ng/ml) we compared the survival probabilities between the two groups. Advanced cases were 17 times (95% CI: 12-24) more likely to die from their disease compared to cases with a localized cancer (Table 3, Figure 5).

Table 3: Prostate cancer-specific survival for CAPS patients diagnosed with a localized vs. advanced disease.

<table>
<thead>
<tr>
<th></th>
<th>Localized Cancer (%)</th>
<th>Advanced Cancer (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>1690 (93)</td>
<td>895 (66)</td>
</tr>
<tr>
<td>Deceased from other events</td>
<td>87 (5)</td>
<td>91 (7)</td>
</tr>
<tr>
<td>Deceased from prostate cancer</td>
<td>34 (2)</td>
<td>364 (27)</td>
</tr>
</tbody>
</table>

Figure 5: Prostate cancer-specific survival for CAPS patients diagnosed with a localized vs. advanced disease.

3.1.2 Prostate cancer families (Paper I)

Swedish families with multiple prostate cancer cases have been collected at the Department of Radiation Sciences, Umeå University, since 1995 (151). Ascertainment of families has been mainly based on referrals by urologists and oncologists throughout Sweden. We confirmed all prostate cancer diagnoses through the National Cancer Register. Almost all the families fulfilled the Carter criteria for HPC.
as described above. For CDH1 analysis, DNA was available for a total of 81 families, comprising 157 men with a FPC or HPC.

3.2 Genotyping Methods

3.2.1 DASH (Paper I)
We used DASH (Dynamic Allele Specific Hybridization) (152) to genotype CDH1 variants. The main principle of DASH is to steadily increase the temperature of a duplex formed between PCR amplified target DNA and an oligonucleotide probe specific to the wild-type allele. The duplex interacts with a double strand–specific intercalating dye which emits fluorescence proportional to the amount of duplex present. The fluorescence signal is monitored as temperature increases and as the duplex denaturates a rapid fall in fluorescence will be seen. Genotype discriminating is then based on melting temperature as a mismatch to the wild-type allele at the SNP site between the probe and the target DNA sequence will result in a lower melting temperature. Details about specific PCR and genotyping conditions are found in paper I.

3.2.2 SEQUENOM (Paper II-V)
Sequenom’s MassARRAY technology uses an allele-specific extension reaction. An “extension” primer anneals to the polymorphic site and is extended. Depending on which SNP allele is present, the primer will extend one or two bases creating a primer extension product. The product mass is then measured to determine the length of the extension product and thereby genotype.

In paper II, we used the MassArray system (SEQUENOM, Inc. San Diego, California) for SNP genotyping. GSTT1 and GSTM1 deletion genotypes were generated through scoring the two alleles (wild type allele and deletion allele) of PCR products on 2% of agarose gel. GSTM3, MSR1 (rs11274081), MSR1 (rs3036811), PGK1, NCOA1, IGF1 and VDR (poly (A) microsatellite) were genotyped using fragment analysis. Details about specific PCR and genotyping conditions are found in paper II.

In order to genotype the CAG repeat of AR, PCR products were supplemented with the internal size standard GS500-LIZ and then separated and detected on an Applied Biosystems model 3730 DNA Analyzer. Alleles were
automatically called using DAC, an allele-calling program developed at deCODE genetics Inc. (153). To determine the absolute number of CAG repeats DNA samples from 188 males were sequenced.

In paper III-V we used matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (154). Genotyping calls were based on peak identification from the mass spectra using the SpectroTYPER RT 2.0 software (Sequenom Inc., San Diego, California, USA). Some SNPs (paper III-V) were genotyped using the iPLEX system which allows for multiplexing of assays up to 29 SNPs. Details about specific PCR and genotyping conditions are found in paper III-V.

3.2.3 Quality Control

For all studies we used blinded duplicate samples to assure high quality of genotyping. More than 100 individuals were genotyped twice in each study and concordance rate between genotype calls were estimated for each SNP. Overall, all studies had a concordance rate >99.9% (range: 99.92-99.98). If a duplicate sample showed dissimilarities for several SNPs it was regarded as contaminated and removed from all analyses. We also assessed Hardy Weinberg equilibrium (HWE) among the controls to assure an adequate genotypic distribution in the population. If a SNP showed strong deviation from HWE (i.e. $P<0.01$) it was removed from further analysis.

3.3 SNP Selection

3.3.1 E-cadherin (Paper I)

CDH1 is located on chromosome 16q22.1 and covers approximately 98 kb. It consists of 16 exons and is divided chiefly into two haplotype blocks (Figure 6). The CDH1 target region included 10 kb of the promoter region and reached 5 kb downstream of the 3’UTR. A total of 23 SNPs were initially selected from public available databases and genotyped in 94 CAPS controls. We obtained successful results for eleven SNPs. A total of six SNPs that captured >95% of the haplotypic variation in CDH1 were selected as tagSNPs. We also included the rs16260 promoter SNP, previously associated with prostate cancer risk, resulting in a total of seven tagSNPs genotyped in CAPS1 and prostate cancer families.
3.3.2 Replication Study (Paper II)

We searched the public library PubMed to identify all reported associations between a polymorphism and prostate cancer risk. Searches lasted until March 1st, 2004. Search terms used were as follows: (prostate, cancer, polymorphisms), (prostate, cancer, association, genetic), (prostate, cancer, SNP), (prostate, cancer, sequence, variants), (prostate, cancer, association), (prostate, cancer, microsatellite). We limited our search to include only studies published in English. An association was considered significant if $P \leq 0.05$ or if the confidence interval did not include 1. A total of 79 polymorphisms were identified. A schematic view of the selection process is presented in Figure 7.

To increase validity of included studies we did not consider associations in a study population with less than 100 cases and 100 controls. Eleven polymorphisms did not fulfill this criterion and was therefore excluded. For different reasons (Figure 7) yet eight polymorphisms were excluded leaving 60 polymorphisms selected for genotyping.
Genetic variants reported to be associated with prostate cancer risk. Polymorphisms were identified by searching the electronic library PubMed until March 1\textsuperscript{st}, 2004. The combinations of words used were: (prostate, cancer, polymorphisms), (prostate, cancer, association, genetic), (prostate, cancer, SNP), (prostate cancer, sequence, variants), (prostate, cancer, association), (prostate, cancer, microsatellite).

\textbf{Figure 7}: Systematic scheme of polymorphism selection in \textit{paper II}
3.3.3 Androgen pathway genes (Paper III and IV)

3.3.3.1 Androgen receptor

AR is located on Xq12 and consists of eight exons, covering approximately 180 kb of the genome. AR harbours within one single haplotype block showing strong LD (Figure 8). Our target region for SNP selection included three kb of the promoter region, all exons, introns and eight kb of 3’UTR. We genotyped 52 initially selected SNPs in 94 unselected controls from the CAPS study. As a result, only 15 SNPs were successfully genotyped and polymorphic in CAPS. Haplotypes were inferred and four SNPs that captured >95% of the haplotypic variation were chosen as tagSNPs. Subsequent comparison with the haplotypic distribution from HapMap prompted us to add two more SNPs to assure satisfactory coverage of the total genetic variation.

Figure 8: LD structure of AR.
3.3.3.2 CYP17
CYP17 is located on 10q24.32 and covers 7 kb. It consists of eight exons and harbours within one single haplotype block that captures 14 kb of the promoter region, all exons and introns and nine kb of the 3'UTR (Figure 9). We downloaded HapMap data for the haplotype block of interest and identified 21 SNPs with a MAF >5%. Out of these 21 SNPs, seven were selected as tagSNPs to capture at least 95% of the common genetic variation.

![Figure 9: LD structure of CYP17.](image)

3.3.3.3 SRD5A2
SRD5A2 stretches 56 kb over 2p23.1. It consists of five exons and is divided into two distinct haplotype blocks (Figure 10). The two haplotype blocks constituted our target region and corresponded to a region reaching 54 kb upstream from the transcription start site to 2 kb of the 3'UTR. We based our SNP selection on HapMap data. We identified 32 SNPs with a MAF > 5%. Each block was separately tagged and to ensure satisfactory coverage we selected additional SNPs to fill the gaps between blocks. Eleven tagSNPs were needed to capture >95% of the common genetic variation in the region.
3.3.4 ERG (Paper V)

ERG is located on 21q22.2 and covers 280 kb. The gene fusion between ERG and TMPRSS2 on 22q22.2 takes place in the 5'UTR of ERG. Based on this information our primary region of interest was the promoter. We based our SNP selection on HapMap data. We only included SNPs with a MAF >5%. By including complete haplotype blocks our target region spanned 121 kb upstream and 1.8 kb downstream from ATG. This region included four distinct haplotype blocks which were tagged separately (Figure 11). tagSNPs were selected to capture at least 95% of the haplotype diversity. In addition, we filled the gap between the blocks by tagging the whole region and chose possible additional SNPs needed to ensure a total coverage of at least 95% of the common genetic variation. In total, 23 tagSNPs were selected.
Figure 11: LD structure of ERG.

3.4 Statistical methods

3.4.1 Haplotype tagging methods (Paper I,III-V)

To select tagSNPs for CDH1 in paper I and AR in paper III and IV, we used the htSNP2 software (155) as implemented in STATA version 8.0 (156). Haplotypes were inferred using a Bayesian approach as implemented in PHASE (157). htSNP2 uses a regression approach to search for an optimal subset of SNPs that maximizes the percentage of haplotypic variation explained, measured by the coefficient of determination. For paper III-V, we used the tagSNPs software (158) to identify the optimal set of tagSNPs. tagSNPs uses a partition-ligation EM algorithm which estimates the number of copies of a specific haplotype (0, 1 or 2) for each subject given their genotypes (so called “dosage”). The squared correlation ($R^2$) between the “true” and the predicted dosage (i.e. the dosage estimated from a subset of SNPs) is then estimated for different SNPs. The SNPs that give the highest $R^2$ are then selected as tagSNPs. We set a limit of $R^2$ to exceed 0.95.

3.4.2 Hardy Weinberg equilibrium (Paper I-V)

We tested all autosomal SNPs for HWE by using a simulation method as implemented in the GENETICS package for the publicly available software R (159). For all tests 10,000 permutations were run.
3.4.3 Association analysis (Paper I-III, V)

3.4.3.1 Polymorphism analysis (Paper I-III, V)
We tested for association between polymorphisms and prostate cancer risk using a covariate corresponding to number of rare alleles (0, 1 or 2) based on unconditional logistic regression. We adjusted for matching between cases and controls in CAPS by including indicator variables representing all combinations between age-group (five years of interval) and geographic region (two groups) in the logistic regression model. To account for dependence between genotypes among relatives in Paper I, we adjusted our analysis using a Huber/White/sandwich method which provides a robust estimation of confidence intervals (160). All analyses were performed in R and STATA.

3.4.3.2 Haplotype analysis (Paper I,III,V)
Haplotypic effects on prostate cancer risk were tested using the HAPLO STATS (161) package as implemented in R. HAPLO STATS measures association between phenotype and haplotypes when phase is ambiguous by using a score test based on a generalized linear model. An iterative method allows simultaneous estimation of haplotype phase and association by using estimated haplotype probabilities for each subject as weights in a regression model. The posterior probabilities for haplotype assignment are updated based on the estimated regression coefficients. Both global and haplotype-specific tests for association were performed and empirical p-values were derived by randomly permuting marker phenotypes. Precision criteria for the p-values were set to a sample standard error of one fourth of the estimated p-value but at least 1,000 permutations were run for each simulation. All haplotype analyses were adjusted for age and geographical region as described above. To calculate haplotype-specific odds ratios we estimated the “dosage” for each subject using the tagSNPS software and included it as a covariate in a logistic regression.

3.4.4 Transmission/Disequilibrium testing (Paper I)
In paper I, we tested if CDH1 alleles were transmitted more often than expected by chance to affected individuals by using a family-based transmission test as implemented in TRANSMIT (162). TRANSMIT uses a robust variance estimator that makes it possible to consider more than one affected offspring in each family. This
method has proven efficient in situations with a high proportion of missing genotypes among parents.

3.4.5 Survival analysis (Paper IV and V)

Time of follow-up for prostate cancer cases in CAPS was calculated from date of diagnosis to date of death or last follow-up (March 1st, 2007). Censoring occurred if a patient was deceased from another cause than prostate cancer or if he was still alive at end of follow-up.

3.4.5.1 SNP analysis (Paper IV and V)

To test for association between prostate cancer-specific survival and a SNP we used a likelihood ratio test of a covariate equal to the number of rare alleles (0, 1 or 2) based on the Cox proportional hazards model. The proportional hazards assumption was tested using Schoenfeld residuals. All analyses were performed in R and STATA.

3.4.5.2 Haplotype analysis (Paper IV and V)

Haplotypic effects on prostate cancer-specific survival were tested with the THESIAS software (163). THESIAS simultaneously estimates haplotype frequencies and haplotypic effects using a stochastic-EM algorithm for likelihood maximization and then assesses association between haplotypes and survival with a standard Cox proportional hazards formulation (164). Simultaneous estimation of haplotype frequencies and haplotype effects is expected to be more efficient in parameter estimation (164). Hazard ratios and corresponding confidence intervals were estimated for each haplotype by comparison to a reference haplotype chosen as the most frequent one. A likelihood ratio test was used as a global test of association between haplotypes and prostate cancer-specific death.

3.4.6 Adjustment for multiple testing through permutation (Paper I-III,V)

We obtained adjusted p-values for each SNP from the empirical distribution of all maximum test statistics based on 10,000 generated data sets. We also estimated the probability of observing at least n significances under the null hypothesis, based on the test statistics from each replica.

3.4.7 Population attributable risk (Paper III)

In order to estimate the impact of the ‘GGAAGA’ haplotype in AR we calculated the population attributable risk (PAR) which estimates the proportion of disease in the
study population that is attributable to a given exposure. PAR was estimated by maximum likelihood estimation as described in (165).
4. Results and comments

4.1 Paper I

The E-cadherin Study

In order to assess association between CDH1 variation and prostate cancer risk we genotyped seven SNPs in CAPS1 and in a family-based material. We replicated earlier findings between rs16260 and prostate cancer risk for cases with a positive family history of prostate cancer (FH+). Carriers of the variant ‘A’ allele exhibited a significant risk increase for prostate cancer ($P$ for trend=0.003). Genotype-specific risk estimates were essentially the same as observed in our initial analysis of rs16260 (112). Merging data from the two studies revealed a risk increase of 47% (95% CI: 1.1-2.0, Table 7) for heterozygous carriers and a 2.6-fold (95% CI: 1.6-4.3) risk increase for homozygous ‘A’ allele carriers compared to homozygous ‘C’ allele carriers.

Table 7: Association between prostate cancer risk and CDH1 promoter SNP rs16260.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Number of subjects (%)</th>
<th>Odds ratio (95 % CI)</th>
<th>Overall P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Controls FH+ cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C/C</td>
<td>397 (53) 153 (42)</td>
<td>1.00 (Ref)</td>
<td>0.0001</td>
</tr>
<tr>
<td>C/A</td>
<td>305 (40) 168 (46)</td>
<td>1.47 (1.1-2.0)</td>
<td></td>
</tr>
<tr>
<td>A/A</td>
<td>50 (7) 46 (12)</td>
<td>2.61 (1.6-4.3)</td>
<td></td>
</tr>
</tbody>
</table>

1 Odds ratios are adjusted for age and geographical region.

Similar association was found for rs4783681 which was strongly correlated with rs16260. Two SNPs were nominally associated with sporadic prostate cancer (rs2010724, $P$=0.02 and rs1801026, $P$=0.04) but these did not reach statistical significance after adjustment for multiple testing. Overall, the haplotypic distribution differed between FH+ cases and controls ($P$ _global_=0.05). Specifically, the ‘AGTGGTC’ haplotype was more common among FH+ cases than controls (31% vs. 25%, $P$=0.004) corresponding to an OR of 1.40 (95% CI: 1.1-1.8).

We used family based tests to investigate if CDH1 variants were inherited to an affected offspring in a larger extent than expected. We had access to 123 families comprising 340 prostate cancer cases and 464 unaffected relatives. DNA was available for an average of 2.1 affected cases and 1.5 unaffected relatives within each family. The variant ‘A’ allele of rs16260 was transmitted in a greater extent than
expected to affected offspring \( (P=0.02) \). In addition, excess transmissions were observed for three other SNPs (rs4783681, rs1125557 and rs2276329). There was overall a distortion in haplotype transmission \( (P_{\text{global}}=0.01) \), mainly attributable to the ‘AGTGGTC’ haplotype that was significantly over-transmitted from parent to affected offspring \( (P=0.02) \).

The agreement between population- and family-based results in this study promotes CDH1 as a prostate cancer susceptibility gene for hereditary prostate cancer. Our lack of findings in the sporadic population suggests that multiple loci, including CDH1, are necessary for prostate cancer development. Due to the random assignment of alleles from parent to offspring, the CDH1 association signal will be attenuated in the general population as the majority will not carry all loci required for disease development. Within families, however, alleles are inherited together making identification of a single locus included in complex patterns with several genetic variants more efficient.

The only haplotype ‘AGTGGTC’ harbouring the rs16260 risk allele was also associated with prostate cancer having essentially the same risk estimates as rs16260. An American study identified a risk excess for carriers of this haplotype in US Caucasians but not in other populations (114). This is in agreement with other studies as the majority reports significant association for rs16260 in Caucasians but not for other ethnicities (110-115). This suggests that rs16260 is a genetic determinant of prostate cancer in men of Caucasian origin but not in other ethnicities (166).

**4.2 Paper II**

*The Replication Study*

In all, 46 selected polymorphisms previously reported to alter prostate cancer risk were assessed for association in CAPS1. The majority of polymorphisms were located in genes involved in the androgen and xenobiotic metabolism (Table 8). Twenty-nine of these polymorphisms are believed to have direct functional consequences by altering the coding sequence, introducing gene splicing or affecting gene expression.
Table 8: Genes tested for association with prostate cancer risk by gene functional group.

<table>
<thead>
<tr>
<th>Gene pathway</th>
<th>Number of genes</th>
<th>Gene</th>
<th>Polymorphisms reported</th>
<th>Associated genes in CAPS</th>
<th>Associated polymorphisms in CAPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xenobiotic metabolites</td>
<td>8</td>
<td>CYP1A1, CYP1B1, GSTM1, GSTM3, GSTP1, GSTT1, NAT1, NAT2</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Androgen metabolites</td>
<td>5</td>
<td>AR, CYP17, CYP19A1, HSD17B3, SRD5A2</td>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>3</td>
<td>CCND1, IGF1, TP53</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Estrogen metabolites</td>
<td>2</td>
<td>ESR1, NCOA3</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Inflammation</td>
<td>2</td>
<td>IL8, MSR1</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Metabolic processes</td>
<td>2</td>
<td>AMACR, PGK1</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Angiogenesis</td>
<td>1</td>
<td>COL18A1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Function Unknown</td>
<td>1</td>
<td>ELAC2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tumor Suppressor</td>
<td>1</td>
<td>LZTS1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DNA repair</td>
<td>1</td>
<td>OGG1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vitamin D</td>
<td>1</td>
<td>VDR</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We replicated six polymorphisms in five different genes (Figure 12); AR (P=0.03), CYP17 (P=0.04), GSTT1 (P=0.006), MSR1 (P=0.009) and SRD5A2 (P=0.02 and P=0.02, respectively). Specifically, carriers of more than 22 repeats of the AR CAG microsatellite had a 24% (95% CI: 1.0-1.5) risk increase to develop prostate cancer. Carriers of two copies of the variant ‘G’ allele of CYP17 rs743572 had a significant risk reduction (OR 0.71, 95% CI: 0.5-1.0). The two SRD5A2 SNPs (rs676033 and rs523349) were in strong LD with each other and showed similar risk effects (OR 1.21, 95% CI: 1.0-1.5) and (OR 1.23, 95% CI: 1.0-1.5) respectively. Missing at least one copy of the GSTT1 gene decreased prostate cancer risk with 23% (95% CI: 0.6-0.9) and the rare MSR1 ‘A’ allele (control frequency 4%) was associated with a risk reduction of 38% (95% CI: 0.4-0.9).
Restricting our analyses to patients having an advanced disease revealed a further accentuated increased risk for carriers of more than 22 CAG repeats in AR and for rs676033 and rs523349 in SRD5A2, implying that these genes may predispose to more advanced prostate cancer. Interestingly, three of five replicated genes (AR, CYP17 and SRD5A2) are involved in the androgen biosynthesis.

The direction of association for all these polymorphisms has shifted between studies suggesting that the causal allele is not yet pin-pointed. Recent relatively large studies failed to replicate the CAG repeat (120, 167), CYP17 (168), SRD5A2 (169) and a recent meta-analysis found no evidence for the MSR1 SNP (170). When adjusting for multiple testing, none of the SNPs remained significant. However, of 46 polymorphisms tested, six (13%) were significant on a 5% level.

Our study illustrates the difficulties in establishing predisposing genetic factors in a complex disease through association analysis. A review of all earlier prostate cancer associations identified for this study revealed that the median number of study subjects was 129 cases and 184 controls. These figures represent far too small sample sizes for adequate analyses. To be able to identify low-penetrant disease susceptibility loci through association analysis large well-conducted studies must be undertaken.

**Figure 12:** Association between earlier associated prostate cancer polymorphisms and prostate cancer risk in CAPS.
4.3 Paper III

The Androgen Pathway Study

In order to follow up on the results from our replication study, we performed a more comprehensive study of the three androgen regulating genes replicated in paper II. Using the approach of haplotype tagging, six AR tagSNPs, six CYP17 tagSNPs and eleven SRD5A2 tagSNPs were successfully genotyped in CAPS.

We identified significant risk reduction for four AR SNPs (Table 9). Our results are in agreement with an Australian study which identified an association between rs6152 and reduced risk of metastatic prostate cancer (171). An American study tested 32 SNPs in AR for association with prostate cancer risk and found that eleven SNPs were associated with a reduced risk for aggressive disease (120). Three CYP17 SNPs located at the 3’ end of the gene decreased prostate cancer risk whereas one SNP located in the promoter region was associated with increased risk. The strongest association was observed for the 3’UTR SNP rs619824, earlier reported to have a protective effect (172). In contrast, we found no evidence that common genetic variation in SRD5A2 affects prostate cancer risk. Of eleven SNPs tested, only one SNP located 28 kb upstream in the promoter was associated with prostate cancer risk on a 5% level.

To account for multiple testing we performed a data simulation by randomly permuting case-control status and then re-evaluated association for each SNP. Based on 10,000 permutations, the only SNPs that remained significant were rs6152, rs7061037 and rs5964607 in AR; however, the probability for observing at least 9 significant associations was estimated to only 0.8%.

We observed an overall difference in AR haplotype frequencies between cases and controls ($P_{global}=0.04$). The most common AR haplotype ‘GGAAGC’ was more prevalent in cases (78%) than in controls (74%) yielding a 25% excess risk of developing prostate cancer (95% CI: 1.1-1.5, $P=0.002$). This was more evident for advanced cases (carrier frequency 80%, OR 1.39, 95% CI: 1.2-1.7, $P=0.0004$). Distribution of CYP17 haplotypes differed between prostate cancer cases and controls ($P_{global}=0.03$). We did not observe any differences in haplotype frequencies between cases and controls at the SRD5A2 locus.
Table 9: Odds ratios and corresponding p-values for SNPs in hormone regulating genes that were associated with prostate cancer risk in CAPS.

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>OR</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>rs17302090</td>
<td>0.75</td>
<td>0.6-1.0</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>rs6152</td>
<td>0.77</td>
<td>0.6-0.9</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>rs7061037</td>
<td>0.77</td>
<td>0.7-0.9</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>rs5964607</td>
<td>0.79</td>
<td>0.7-0.9</td>
<td>0.004</td>
</tr>
<tr>
<td>CYP17</td>
<td>rs2486758</td>
<td>1.15</td>
<td>1.0-1.3</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>rs10883783</td>
<td>0.94</td>
<td>0.8-1.1</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>rs4919683</td>
<td>0.90</td>
<td>0.8-1.0</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>rs619824</td>
<td>0.90</td>
<td>0.8-1.0</td>
<td>0.009</td>
</tr>
<tr>
<td>SRD5A2</td>
<td>rs623419</td>
<td>1.14</td>
<td>1.0-1.3</td>
<td>0.02</td>
</tr>
</tbody>
</table>

We estimated the joint effect of identified risk alleles from all three genes by calculating combined risk estimates for rs6152 in AR, rs619824 in CYP17 and rs623419 in SRD5A2. Compared with the reference group (carriers of zero risk alleles), individuals carrying five risk alleles exhibited a significant risk increase (OR 1.87, 95% CI: 1.0-3.4). For each additional risk allele carried, the risk of developing prostate cancer was increased by 12% (95% CI: 1.1-1.2, \( P = 0.00009 \), Table 10). Similar estimates were observed for advanced cancer (OR 1.13, 95% CI: 1.1-1.2, \( P = 0.0008 \)), and more pronounced for early onset of disease (OR 1.20, 95% CI: 1.1-1.30, \( P = 0.00007 \)). Our results suggest that combined analysis of multiple SNPs may strengthen otherwise nominally associations with individual SNPs. This is in agreement with the hypothesis of prostate cancer being a multigenic disease. Combining moderate risk effects may ultimately result in identification of individuals with a noteworthy risk increase. There was no statistical correlation between the genotypes on a multiplicative scale (\( P = 0.61 \)). More likely, our observations are a result of summing up individual main effects.

We also investigated the relationship between AR SNPs and the CAG repeat in exon 1. We defined a long allele as \( \geq 22 \) repeats. We observed a strong correlation between haplotypes and number of CAG repeats (Figure 13). Carriers of the long allele have a low diversity of haplotypes and are mainly carrying the ‘GGAAGC’ and ‘GGAAGT’ haplotypes. These results question the causality of the CAG repeat. Probably, earlier findings have reflected the strong LD structure with adjacent SNPs that appear to have a more significant role in prostate cancer aetiology.
### Table 10: Combined effects of risk alleles for rs6152, rs619824 and rs623419.

<table>
<thead>
<tr>
<th>Population</th>
<th>Number of risk alleles</th>
<th>Cases (%)</th>
<th>Controls (%)</th>
<th>OR</th>
<th>95% CI</th>
<th>P (trend)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Cases</strong></td>
<td></td>
<td>0</td>
<td>34 (1.2)</td>
<td>29 (1.7)</td>
<td>1.00</td>
<td>REF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>318 (11.3)</td>
<td>238 (14.0)</td>
<td>1.18</td>
<td>0.7-2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>835 (29.6)</td>
<td>550 (32.3)</td>
<td>1.33</td>
<td>0.8-2.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1,041 (36.8)</td>
<td>564 (33.1)</td>
<td>1.64</td>
<td>1.0-2.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>491 (17.4)</td>
<td>272 (16.0)</td>
<td>1.57</td>
<td>0.9-2.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>107 (3.8)</td>
<td>52 (3.1)</td>
<td>1.87</td>
<td>1.0-3.4</td>
</tr>
<tr>
<td><strong>Advanced Cases</strong></td>
<td></td>
<td>0</td>
<td>15 (1.2)</td>
<td>29 (1.7)</td>
<td>1.00</td>
<td>REF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>145 (11.6)</td>
<td>238 (14.0)</td>
<td>1.20</td>
<td>0.6-2.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>364 (29.1)</td>
<td>550 (32.3)</td>
<td>1.26</td>
<td>0.7-2.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>454 (36.3)</td>
<td>564 (33.1)</td>
<td>1.55</td>
<td>0.8-2.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>218 (17.4)</td>
<td>272 (16.0)</td>
<td>1.53</td>
<td>0.8-3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>55 (4.4)</td>
<td>52 (3.1)</td>
<td>2.13</td>
<td>1.0-4.4</td>
</tr>
<tr>
<td><strong>Young Cases (&lt; 65 years)</strong></td>
<td></td>
<td>0</td>
<td>17 (1.2)</td>
<td>15 (2.2)</td>
<td>1.00</td>
<td>REF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>177 (12.4)</td>
<td>105 (15.2)</td>
<td>1.57</td>
<td>0.8-3.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>407 (28.6)</td>
<td>234 (33.9)</td>
<td>1.61</td>
<td>0.8-3.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>522 (36.7)</td>
<td>220 (31.8)</td>
<td>2.20</td>
<td>1.1-4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>248 (17.4)</td>
<td>106 (15.3)</td>
<td>2.12</td>
<td>1.0-4.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>53 (3.7)</td>
<td>11 (1.6)</td>
<td>4.35</td>
<td>1.7-11.3</td>
</tr>
</tbody>
</table>

**Figure 13:** Relationship between AR haplotypes and the CAG repeat in exon 1 of AR.
4.4 Paper IV

The Androgen Pathway and Survival Study

This study represents the first utilization of follow-up data in CAPS for identification of factors important for disease progression. Twenty-three SNPs, earlier selected to tag AR, CYP17 and SRD5A2, were tested for association with prostate cancer-specific death. We observed overall no association between SNPs and prognosis. To explore if variation within these genes were important for hormonal treatment response we performed sub-analysis on 918 subjects (269 prostate cancer-specific deaths) that received hormonal therapy as their primary treatment. Carriers of the variant ‘A’ allele of the AR rs17302090 SNP had an almost two-fold risk to die from prostate cancer (HR 1.93, 95% CI: 1.2-3.0, P=0.007) compared to non-carriers. In contrast, we found no significant association between rs17302090 SNP and survival among patients who did not receive any hormonal treatment (HR 1.61, 95% CI: 0.5-5.3, P=0.46).

We found overall no association between haplotypes and prognosis for any of the genes. When restricting our analysis to include only hormonally treated patients we noticed that an AR haplotype ‘AAGAGT’ increased the risk of lethal prostate cancer with 30% (95% CI: 1.1-1.7, P=0.009).

This is the first study that investigates the association between common genetic variants in these genes and prostate cancer prognosis. Earlier studies have investigated the AR CAG repeat with PSA recurrence, failure of hormonal response and disease-specific survival as primary end-points. These studies comprise sparse study populations and have shown equivocal results (124-130). We performed a prostate cancer-specific survival analysis on 1,295 patients earlier genotyped for the CAG repeat (171 prostate cancer deaths) and observed no association (P=0.32).

This study proposes an interesting hypothesis that may in part explain the variability in time to recurrence among hormonal treated patients. If treatment response depends on an individual’s genetic background it may be possible to identify those patients that should be offered other alternatives.
4.5 Paper V

The ERG Study

To explore the possible influence of common genetic variants in ERG 5’UTR on prostate cancer risk and progression a total of 21 tagSNPs were successfully genotyped and tested for association in CAPS. We observed no association between individual SNPs or haplotypes and prostate cancer risk. These data suggest that if common genetic variation at the ERG locus has a notable influence on prostate cancer risk, it would be found outside our target region and demonstrate no LD with SNPs and haplotypes examined in this study. Two common SNPs, rs2836626 and rs2836582 located in block four, were associated with prostate cancer-specific death. Carriers of the rs2836626 ‘T’ allele (MAF = 21%) were at increased risk to die from prostate cancer (HR 1.28, 95% CI: 1.1-1.5, \( P=0.009 \)) whereas carriers of the rare ‘T’ allele (MAF=23%) of rs2836582 had a significant better prognosis (HR 0.80, 95% CI: 0.7-1.0, \( P=0.02 \)). Multivariate Cox regression analysis adjusted for TNM-stage, Gleason sum and PSA level at diagnosis did not alter these findings. Carriers of the rs2836626 ‘T’ allele were diagnosed with a significant higher TNM-stage than non-carriers (\( P=0.009 \)) suggesting that genetic variation in this area may affect tumor stage and thereby alter the risk of dying from prostate cancer. Indeed, in survival analysis only adjusted for TNM-stage the association between rs2836626 and prostate cancer-specific survival disappeared. No other SNP was associated with prostate cancer-specific death. After adjustment for multiple testing no SNP remained significant.

We observed borderline association between haplotypes and prostate cancer-specific death in block four (\( P_{\text{global}}=0.06 \)). Specifically ‘CTCGTATG’ carriers had a 36% increased risk to die from prostate cancer compared to the most common haplotype (95% CI: 1.1-1.7, \( P=0.006 \)). The variant ‘T’ allele of rs2836626 is only present at the ‘CTCGTATG’ haplotype suggesting that this haplotype harbours a yet untyped causal marker in strong linkage disequilibrium (LD) with rs2836626.

This is the first study to evaluate if polymorphisms located upstream of ERG are involved in prostate cancer development and progression. Block four is located 100 kb upstream of ERG and spans a region of 57 kb. Because we did not have access to tumor samples we could not correlate SNPs and haplotypes with TMPRSS2:ERG fusion status. Another limitation is the restricted genetic region
investigated. We can not exclude that genetic variation outside this region plays a significant role in prostate cancer.
5. Discussion

5.1 Evidence of genetic predisposition to prostate cancer (Paper I-III, V)

Although there is large body of evidence supporting the importance of genetic factors in prostate cancer development the success in identifying the relevant loci has been limited. This thesis aims to investigate if polymorphisms in specific candidate genes contribute to prostate cancer development and progression.

In paper I, we performed a comprehensive investigation of the E-cadherin gene, a tumor suppressor gene earlier implicated in prostate cancer. We replicated a common promoter SNP (rs16260) known to alter the transcriptional activity of the gene. Our data suggests that E-cadherin variation explains a fraction of the hereditary prostate cancer cases in Sweden. Interestingly, rs16260 was borderline associated with a reduced risk of dying from prostate cancer (HR=0.84, 95% CI: 0.7-1.0, P=0.05) in CAPS.

In paper II, we set out to evaluate the state of the art of prostate cancer genetic epidemiology in 2004. We identified all reported associations between polymorphisms and prostate cancer risk and assessed them for association in CAPS1. Six out of 46 (13%) polymorphisms were associated with prostate cancer in CAPS, a higher number than expected by chance. Interestingly, three of the replicated genes (AR, CYP17 and SRD5A2) are involved in the biosynthesis of androgens, implying the importance of this pathway in prostate cancer. Summing up evidence from association studies of these genes reveals inconsistency and contradictions. This study reflects the difficulties with replication in association studies. We set out to assess all polymorphisms reported to associate with prostate cancer in a minimum study population of 100 cases and 100 controls. We excluded one polymorphism due to inadequate bioinformatics available at that time emphasizing the importance of bioinformatics, especially as the number of known SNPs increases exponentially. The introduction of rs numbers as identifiers and the public availability of SNP databases will prove an important step to facilitate reproducible studies. We did not manage to design assays or obtain reliable genotypes for 14 of the polymorphisms. As our genotyping quality control did not indicate any internal problems with our genotyping method these polymorphisms may be located in “problematic regions” in the genome, i.e. regions with repetitive sequences, copy number variation,
adjacent SNPs and so forth. If so, it is difficult to interpret earlier findings, especially as quality control for genotyping was sparsely reported in the original studies.

The importance of androgen regulating genes was further supported by our results in paper III. Several SNPs in both AR and CYP17 showed association with risk and combining the genes resulted in noteworthy risk excess (two-fold for individuals carrying all risk alleles). The risk excess due to combination of alleles should be attributed to the sum of individual main effects from each SNP as we observed no statistical interaction between the genotypes (\(P=0.61\)). Finally, the lack of association in the upstream region of ERG (paper V) indicates that it is not involved in prostate cancer development.

5.2 Evidence of genetic contribution to prostate cancer progression (Paper IV and V)

The importance of inherited genetic variation in prostate cancer progression is a fairly unexplored area. Although we understand several hallmarks of cancer progression including angiogenesis, telomerase activity and aberrant apoptosis programs there are still many factors that are unknown. It has been shown that the survival of women with familial breast cancer is predicted by the prognosis of her first-degree relative with breast cancer (173) and similar results have been found in prostate cancer families (personal communication, Linda Lindström) suggesting that inherited variants indeed may contribute to prostate cancer progression.

There is an urgent need to identify better prognostic factors for prostate cancer. Ideally, this would result in three important things; 1) To better identify cancers that need an aggressive treatment at an early stage and consequently decrease death rates. 2) To avoid over-treatment of latent cancers and consequently reduce treatment side effects that may have a serious impact on quality of life. 3) To develop more targeted and effective treatments.

Since androgens are essential in the natural history of prostate cancer we had a strong priori to believe that hormone regulating genes would have an impact on survival. Even though we did not find any overall association with prostate cancer-specific survival in paper IV it is intriguing to speculate about the possible role of AR in response to hormonal therapy. If some individuals are predisposed to have a poor treatment response, complementary or alternative treatments should be offered. In paper V, we observed association with prostate cancer prognosis in a
region located approximately 100 kb from ERG. Considering the lack of information about this region and that no other study to date has reported similar analyses, it is difficult to draw any conclusions from our results.

5.3 Genetic epidemiology and association studies – Design, strengths and limitations

Association studies have been proposed as a powerful tool to identify common genetic variants involved in complex polygenic diseases. So far, results have been mixed. There are several explanations why association studies have not proven as effective as originally proposed. In prostate cancer epidemiology particularly, several methodological issues have to be overcome in order to design a powerful case-control study. Prostate cancer is a heterogeneous disease and the distinction between cases and controls may be vague. PSA screening is at present not generally implemented in Sweden leading to a high proportion of symptomatic cancers in our study population (174). Nevertheless, an increasing awareness about prostate cancer has led to a more frequent occurrence of PSA testing in the asymptomatic population. The wide phenotypic heterogeneity complicates the classification of study subjects resulting in reduced statistical power (175). To overcome this issue, researchers have suggested different approaches for control selection. For example, several groups have utilized a set of “super controls” with a PSA serum level below four and a normal digit rectal exam (176,177). Although this approach minimizes the possible misclassification of “true” cases as controls it results in a highly selective control group which may not be representative for the general population. The wide variability in disease aggressiveness poses yet a problem. The genes that influence susceptibility to a less aggressive cancer may differ from those that influence the risk of aggressive tumors. Failure of differentiation between these subgroups may reduce statistical power in the analysis and complicate replications in other populations with different phenotypic characteristics (101).

Generally speaking, genetic epidemiology is spared from many of the common pitfalls that arise in classical epidemiological studies. Selection bias is in general not an issue in a genetic association study as it would require specific genotypes to be associated with the ability to participate in the study. Although it seems unlikely that the willingness to participate would be correlated to a genetic marker this could be the case if individuals in prostate cancer families were more
likely to participate. If so, we could expect an enrichment of susceptibility allele carriers in our population as the probability of carrying a disease allele would be higher for those individuals compared to the general population. If the individual is classified as a control, the association will be biased towards the null and if the individual is a case, the association will be biased from the null. However, in order for this scenario to have a notable impact on the estimate, the polymorphism has to have a significant impact on disease risk.

A second common problem in epidemiology is confounding, i.e. when an unmeasured factor drives the association and introduces biased risk estimates between exposure and outcome. In order for a factor to be a true confounder it must precede the exposure. If this is not the case, it should be regarded as an intermediate factor and should not be adjusted for in the analysis. As we are given our genome at conception no environmental factors can precede the exposure of interest in genetic association studies. Possibly, exposure before conception may affect germline DNA (e.g. radiation). The only profound confounder in genetic epidemiology is population stratification (178). This phenomenon occurs if different subpopulations are represented in a study, i.e. admixture. If a genetic marker is more prevalent in a subpopulation that has a high incidence of disease, a spurious association will arise even though the marker itself is not causing disease. To overcome this issue it is important to either stratify the analysis for ethnicity or in other ways adjust for it.

An important concept in both classical and genetic epidemiology is effect modifiers. Effect modifiers alter the magnitude of the risk associated with a given exposure. This would be the case in gene-environment interaction where the genetic marker may act as an effect modifier for a certain exposure. An example is aspirin use, UGT1A6 variation and the risk of colorectal adenoma. Wild-type carriers do not benefit from regular aspirin use whereas carriers of the variant allele have a decreased risk (179, 180).

When interpreting the results from a genetic association study all issues mentioned above must be considered. However, it is important to stress that an association study is not itself sufficient to establish causality. Despite strict control of misclassification, selection bias and confounding, there is always a risk of random errors. The best way to counteract these is to conduct well-powered studies relatively insensitive to statistical fluctuations, but the problem of random errors can never be completely removed. In addition, despite careful sampling of the study population
there is always a possibility that the population in general is not correctly represented.

5.4 CAPS – Design, strengths and limitations

The CAPS study was originally designed as a classical epidemiological case-control study with dietary factors as the main exposure. Along with this, cases and controls were matched according to age and geographical region. Although age is the strongest risk factor for prostate cancer the only situation where age could be a confounder in genetic epidemiology is if the distribution of the assessed genetic variant is skewed among age categories in the population. The geographical region would be a more prominent concern. Subpopulations with different genetic background could exist leading to spurious findings. The incidence of prostate cancer in Sweden 2002 varied between different counties with Uppsala having the highest incidence (203/100,000) and Gotland the lowest (83/100,000) (54). If these regions also show genetic diversity, false association signals may appear. To assess the probability of population stratification in CAPS we tested approximately 8,000 SNPs for association and utilized both unadjusted and adjusted analyses (200 randomly selected controls and 200 randomly selected unrelated familial cases). These SNPs were selected on the basis of their significance for inflammation and thereby plausible determinants of prostate cancer development. As can be seen in Figure 14, there were no difference between adjusted and unadjusted analysis suggesting that there is no indication of population stratification in our population.
Another important issue in case-control studies is the participation rate in cases and controls. In CAPS we obtained blood samples from 81% of all invited cases and 58% of the invited controls. Although we identified incident cases there is still a possibility that a selection is introduced. Men diagnosed with a severe disease may not be able to participate because of their health condition. In some cases, the patient was deceased at time for recruitment. Consequently, CAPS may be selected towards less severe cancers. The low participation rate among controls results in a distribution of 1.6 cases for each control. Not only does the low participation rate among controls reduce statistical power in our analyses, it also raises the question about selection bias. For a questionnaire-based epidemiological study this could indeed be a prominent issue. If the willingness to participate depends on the ability to answer questions we could expect a selection bias. In order to evaluate the possible impact of selection bias in CAPS we compared characteristics of participants who completed both the questionnaire and contributed with a blood sample to those who only completed the questionnaire. Among both cases and controls, we observed no differences regarding baseline characteristics arguing that if a selection bias in CAPS exists, it is of negligible magnitude (60).

The CAPS study is a large case-control study with nearly 3,000 cases and 2,000 controls. Consequently, given that we manage to pinpoint the causal allele, CAPS minimizes the possibility of type II errors (i.e. the probability of failing to reject
Genetic Variation and Prostate Cancer

the null hypothesis when the alternative hypothesis is true). Assuming an additive inheritance model with an odds ratio of 1.20 for heterozygous carriers and 1.44 for homozygous carriers CAPS has 80% power to detect a SNP with an allele frequency of 11% (Figure 15).

![Figure 15: Statistical power in CAPS as a function of allele frequency among the controls. The calculations are based on 2,965 cases and 1,823 controls. An additive inheritance model with an odds ratio of 1.2 for heterozygous carriers and 1.44 for homozygous carriers was assumed.](image)

As previously mentioned, studies of prostate cancer aetiology suffer from a phenotypic heterogeneity, both within and between studies. The introduction of PSA screening worldwide has dramatically changed the clinical characteristics of the disease population. Although PSA screening is not generally introduced in Sweden there has been a shift towards more asymptomatic localized tumors. A third of the tumors in CAPS are T1c tumors, i.e. they are identified through a needle biopsy prompted by an elevated serum PSA level. At March 1st, 2007; only 20 of 1,055 (2%) T1c patients had deceased from prostate cancer, indicating that patients identified through elevated PSA are not likely to progress and develop metastases during a five year period.

Phenotype classification is a difficult task in prostate cancer as the disease displays a wide heterogeneity. All cases in CAPS had a confirmed prostate cancer reported to the National Cancer Register in Sweden. When recruiting controls, we did not control for PSA levels or perform any other actions to estimate their probability of having a prostate tumor. Considering the commonness and latent
nature of prostate cancer it is plausible that a fraction of CAPS controls carries small unidentified tumors in the prostate. Subsequent follow-up revealed 29 controls diagnosed with prostate cancer since inclusion (last follow-up February 15th, 2005) and 320 controls had a PSA level ≥ 4 ng/ml at time of blood draw in subsequent serum analysis of PSA levels. To select controls according to PSA levels at a given cut-off (for example four) would not only result in power loss due to smaller sample size, it would also introduce a selected control group not representative for the general population. In addition, exclusion of initially selected controls may introduce bias in subsequent analysis. We have therefore chosen not to exclude any controls in our analyses even though the possible misclassification of “true” cases as controls will result in a slight dilution of any true association.

We determined the cause of death for all patients included in CAPS through linkage to the Cause of Death Register provided by the National Board of Health and Welfare. In March 15th 2007, 398 cases in CAPS had deceased due to prostate cancer. As we did not have access to coded information about cause of death for individuals who deceased after December 31st, 2004 we acquired copies of their death certificates and let an experienced oncologist review them in order to establish cause of death. To evaluate our agreement with the register we compared the reported cause of death from the register with our own classification based on death certificates for all men who deceased during 2004. We observed a 95% agreement between our own classification and the register, making us confident with our death certificate reviews.

One possible limitation with the Cause of Death Register is that is relies on coded information based on death certificates. Certificates may be sparse and vague making it difficult to determine if prostate cancer is the primary cause of death. A straightforward approach to evaluate the register is to compare the cause-specific survival with relative survival (i.e. the excess mortality in the population due to a specific disease). Relative survival can be defined as the difference between observed mortality and expected mortality in the general population. A recent study originating from Sweden followed 8,887 men diagnosed with prostate cancer for 15 years and compared relative survival with prostate cancer-specific death. They observed a strong concordance between relative and cause-specific death (Figure 16), (181).
Figure 16: The relationship between relative survival and prostate cancer-specific death in Sweden based on 8,887 men diagnosed with prostate cancer between 1987 and 1999. Adapted from Aus et. al, Cancer, 2005 (181).

Another possible limitation with the Cause of Death Register in Sweden is their loss of follow-up for emigrated individuals. To date, we have not collected information about emigration status for individuals in CAPS. That is, if a case patient has emigrated from Sweden he will be classified as alive in our analyses, possibly introducing biased risk estimates. However, it is not likely that this will have an impact on our analyses as the emigration rate in Sweden is low (0.5% in 2006) (182).

5.5 Study design and execution - Molecular and statistical methods
Throughout these studies we have had an extensive quality control to assure reliable genotyping results. To avoid possible systematic misclassification of genotypes between cases and controls DNAs were arranged randomly, and the lab was blinded to case-control status. For all studies, we obtained high success rate (>90%) and high concordance between blinded duplicate samples (>99.9%). We experienced some difficulties with genotyping in the replication study and lost 14 polymorphisms initially anticipated to be included. This was an unexpectedly high number as we had no indication of internal problems with our genotyping methods.

We chose to use a haplotype tagging based approach in all studies except in paper II, which was based on earlier reported specific polymorphisms. We
preferred haplotype tagging as it requires no prior beliefs about specific SNPs. Although coding SNPs might be more appealing due to their known function, several regulatory variants that may be of importance harbour in untranslated DNA. Clearly, haplotype tagging provides a more satisfying coverage of the genetic variation in a region compared to single SNP studies. However, haplotypes do not explain the total genetic variation and there is still a chance that we have failed to identify a true association, especially if the causal allele is rare. In paper III, IV and V, we used HapMap data to select our tagSNPs. Although it has been shown that HapMap data is sufficient to capture common variation in the genome it is important to stress that initial SNP density, MAF, study population and LD structure all have a substantial impact on the ability to capture common variation (25,26).

We addressed the issue of multiple testing through simulation by permuting case-control status and re-evaluating associations. Except for three SNPs in paper III, all adjusted p-values did not reach statistical significance on a 5% level. The problem of multiple testing is hard to tackle. We are aware that genetic association studies suffer from a high number of false positive findings and with increasing number of tests, significant results will eventually appear by chance. This emphasizes the importance of replicating genetic associations in independent populations. On the other hand, as we do not expect large-scale differences in genotype frequencies between cases and controls in a complex disease such as prostate cancer, we do not expect high risk effects. As the number of study participants is not infinite, a modest risk elevation will only provide nominal p-values which probably will fail to retain statistical significance after adjustment for multiple testing. That is, a true association might be disregarded as a random finding after adjustment for multiple testing.
6. Summary and conclusions

In general, single genetic association studies do not lead to conclusions about causality. They constitute an efficient technique to screen the genome for possible susceptibility loci but they do not provide robust scientific evidence. To become a “true” disease locus, consistent replication in several well-powered studies is required. To complete the picture, association studies must be followed by resequencing and functional studies in order to pinpoint the causal allele(s) and to understand the biological mechanisms. Although this may sound discouraging, I believe that implications from genetic association studies will constitute a solid foundation for future efforts.

6.1 Future prospects – Genetic epidemiology of complex diseases

The field of genetic epidemiology has experienced radical changes and achievements during the last three years. In the early days of genetic epidemiology, case-control studies constituted of a single candidate polymorphism genotyped in a small set of subjects. The polymorphism of interest was often known to alter the protein sequence or gene expression. As the knowledge of the human genome increased, researchers started to argue that non-coding elements (known as “junk-DNA”) could also have an impact on gene function. Candidate polymorphism studies were broadened to become candidate gene studies and haplotype analysis was introduced. The HapMap project was launched in 2002 with the aim to characterize a haplotypic map of the genome in order to facilitate candidate gene studies for individual researchers. The idea of candidate gene studies has now been further developed into pathway-driven studies where multiple genes are simultaneously analyzed. Haplotyping is an appealing alternative to single polymorphism studies as it confers a more agnostic approach. Instead of one single hypothesis, the common variation at a specific locus is of interest. However, it has been argued that HapMap only constitutes a temporary substitute since publications of genome-wide studies have begun to emerge and it remains to be seen what role HapMap will have for disease mapping in the future (183).

The recent possibility of genotyping half a million SNPs in thousands of individuals has paved the way towards finding genetic determinants of different traits. Large-scale genotyping is now economically feasible and the high through-put
capacity makes genotyping of large scale data sets a matter of days rather than months. Since December 2006, there has been an explosion of genome-wide association studies for complex diseases resulting in numerous confirmed susceptibility loci (Table 11). We are likely to experience an exponential increase in identified susceptibility genes for complex diseases in the near future. So far, only the strongest associations have been reported and follow-up studies from scans will most likely result in additional loci. Not only will promising loci from individual scans be further evaluated (e.g. chromosome 17q in prostate cancer), but combining data from several scans in the same disease will also result in new susceptibility loci (e.g. type 1 diabetes (35) and type 2 diabetes (184)). Two interesting observations can be made from Table 11. The similarity in identified loci between different scans for the same disease is striking, supporting their role in disease development (although data presentation probably is biased due to previous scan publications). The other interesting observation is that different diseases share the same susceptibility loci. Gudmundsson and colleagues found that a common genetic variant associates with both prostate cancer and type 2 diabetes (37). As several epidemiological studies have demonstrated an inverse relationship between type 2 diabetes and prostate cancer risk (93) the identification of a mutual genetic variant may in part explain that relationship. Another example is chromosome 9p21 that has been shown to be involved in both heart disease and diabetes (33,43).
<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of SNPs</th>
<th>Cases/controls</th>
<th>Chromosomal loci/Gene</th>
<th>MAF</th>
<th>Overall OR</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>266,722</td>
<td>408/400</td>
<td>FGFR2, TNRC9, MAP3K1, LSP1, 8q</td>
<td>48-46</td>
<td>1.07-1.26</td>
<td>(36)</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>528,173</td>
<td>1,145/1,142</td>
<td>FGFR2</td>
<td>39</td>
<td>1.20</td>
<td>(38)</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>316,515</td>
<td>1,453/3,064</td>
<td>8q24</td>
<td>9</td>
<td>1.60</td>
<td>(42)</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>550,000</td>
<td>1,172/1,157</td>
<td>8q24</td>
<td>50</td>
<td>1.26</td>
<td>(41)</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>310,520</td>
<td>1,501/11,290</td>
<td>17q12, 17q24.3</td>
<td>46-49</td>
<td>1.20-1.22</td>
<td>(37)</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>550,163</td>
<td>940/965</td>
<td>8q24.21</td>
<td>47</td>
<td>1.21</td>
<td>(39)</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>99,632</td>
<td>1,257/1,336</td>
<td>8q24</td>
<td>44</td>
<td>1.17</td>
<td>(40)</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>521,000</td>
<td>563/1,146</td>
<td>KIAA0350, PTPN22 INS</td>
<td>28-39</td>
<td>0.65-0.66</td>
<td>(31)</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>500,568</td>
<td>2,000/3,000</td>
<td>1p13, MHC, 12q13, 12q24, 16p13</td>
<td>10-42</td>
<td>1.19-5.49</td>
<td>(185)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>386,731</td>
<td>1,464/1,467</td>
<td>CDKN2A/B, CDKAL1, IGF2BP2</td>
<td>29-83</td>
<td>1.12-1.20</td>
<td>(32)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>315,000</td>
<td>1,161/1,174</td>
<td>CDKN2A/B, CDKAL1, IGF2BP2, SLC30A8, TCF7L2 HHEX, EXT2, LOC387761</td>
<td>30-85</td>
<td>1.08-1.20</td>
<td>(33)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>392,935</td>
<td>686/689</td>
<td></td>
<td></td>
<td>1.14-1.65</td>
<td>(186)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>313,179</td>
<td>1,399/5,275</td>
<td>CDKAL1, SLC30A8</td>
<td>26-67</td>
<td>1.15-1.20</td>
<td>(34)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>500,568</td>
<td>2,000/3,000</td>
<td>6p22, 10q25, 16q12</td>
<td>18-40</td>
<td>1.18-1.36</td>
<td>(185)</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>500,568</td>
<td>2,000/3,000</td>
<td>9p21</td>
<td>47</td>
<td>1.47</td>
<td>(185)</td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>100,000</td>
<td>322/312</td>
<td>9p21</td>
<td>49</td>
<td>1.26</td>
<td>(43)</td>
</tr>
<tr>
<td>Myocardial Infarction</td>
<td>305,953</td>
<td>1,607/6,728</td>
<td>9p21</td>
<td>45-49</td>
<td>1.25-1.28</td>
<td>(187)</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>316,515</td>
<td>550/4,476</td>
<td>4q25</td>
<td>11</td>
<td>1.88</td>
<td>(188)</td>
</tr>
<tr>
<td>Obesity</td>
<td>116,204</td>
<td>694</td>
<td>2q14.1</td>
<td>37</td>
<td>1.22</td>
<td>(189)</td>
</tr>
<tr>
<td>Obesity related traits</td>
<td>362,129</td>
<td>1,412</td>
<td>FTO, PFKP</td>
<td>12-46</td>
<td></td>
<td>(190)</td>
</tr>
<tr>
<td>Inflammatory bowel disease</td>
<td>308,332</td>
<td>567/571</td>
<td>IL23R</td>
<td>7</td>
<td>0.45</td>
<td>(191)</td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>317,503</td>
<td>988/1,007</td>
<td>1p31.2, 2q37.1, 4p13, 10q21.1, 16q12.1, 16q24.1 22q12.3</td>
<td>8-46</td>
<td>1.11-1.52</td>
<td>(192)</td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>500,568</td>
<td>2,000/3,000</td>
<td>1p31.2, 2q37.3p21, p13, 5q33, 10q21, 10q24, 16q12, 18p11</td>
<td>7-48</td>
<td>1.09-1.54</td>
<td>(185)</td>
</tr>
<tr>
<td>Human gallstone disease</td>
<td>&gt;500,000</td>
<td>280/360</td>
<td>ABCG8</td>
<td>8</td>
<td>2.2</td>
<td>(193)</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>310,605</td>
<td>778/1,422</td>
<td>4q27</td>
<td>18</td>
<td>0.63</td>
<td>(194)</td>
</tr>
<tr>
<td>Sporadic amyotrophic lateral sclerosis</td>
<td>766,955</td>
<td>386/542</td>
<td>FLJ10986</td>
<td>32</td>
<td>1.35</td>
<td>(195)</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>500,568</td>
<td>2,000/3,000</td>
<td>1p13, MHC</td>
<td>10-49</td>
<td>1.82-2.36</td>
<td>(185)</td>
</tr>
<tr>
<td>Childhood asthma</td>
<td>317,000</td>
<td>994/1,243</td>
<td>17q21</td>
<td>48</td>
<td>1.41</td>
<td>(196)</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>500,568</td>
<td>2,000/3,000</td>
<td>16p12</td>
<td>28</td>
<td>2.08</td>
<td>(185)</td>
</tr>
<tr>
<td>Restless legs</td>
<td>236,758</td>
<td>401/1,644</td>
<td>MEIS1, BTBD9, MAP2K5</td>
<td>24-33</td>
<td>1.53-1.74</td>
<td>(197)</td>
</tr>
</tbody>
</table>

<sup>1</sup>Initial scan
Todd and colleagues followed up a genome-wide scan in type 1 diabetes and identified four new susceptibility regions (35). To date, ten type 1 diabetes loci have been identified (Figure 17). The distribution of odds ratios for the ten loci illustrates a phenomenon that is likely to occur also in other complex diseases. However, most diseases are unlikely to have a risk allele corresponding to the HLA locus which increases the risk of type 1 diabetes seven-fold. The vast majority of susceptibility alleles will most likely have low or moderate risk effects (OR: 1.1-1.4). The probability of discovering these loci will depend on their allele frequency. No polymorphism described in Figure 17 has an allele frequency below 0.1. Most likely, numerous rare variants with small risk effects exists and these are unlikely to be identified using techniques and study samples available today. Although these variants will be negligible in clinical practice it is important to find them as they may contribute to the understanding of the biological processes causing disease.

![Figure 17: Confirmed susceptibility loci for type 1 diabetes.](image)

Genome-wide association studies are compelling in their agnostic nature. As new susceptibility loci are discovered we will gain novel knowledge in both genetics and biology. However, despite the possibilities that genome-wide scans offer there are some limitations in design and interpretation that need attention. The population
chosen for the initial scan needs to be large enough to have sufficient power to detect true associations. The genetic variants identified from the first line of genome-wide studies are probably the most “easy” to find, considering allele frequency and risk effects. To detect rare variants with low risk effects, initial scans in larger populations will be required. Moreover, the huge amount of data requires development of sophisticated methods to recognize true positives and avoid false negatives. In addition, there are still gaps in the genome that have not been investigated, leaving unexplored regions. This latter issue will not be completely solved until it is financially and technologically feasible to resequence the entire genome on a large-scale basis. Furthermore, we still have limited knowledge of the structure of the human genome. For example, we are only at the beginning of understanding the concept of copy number variation and its possible role in disease development. It is important to stress that there is still a tremendous amount of work ahead. Although a genetic region of interest is identified through a genome-wide association study, additional efforts are needed to pinpoint the specific mutation. In the end, the major challenge will be to understand in what way these genetic variants contribute to disease susceptibility.

Genome-wide association studies undoubtedly constitute an enormous breakthrough in genetic epidemiology. We are at the moment unraveling the genetics behind common diseases such as cancer, heart disease and diabetes. To complete the picture, gene-gene and gene-environment interaction are key issues to address in the near future. In order to achieve successful results in these areas, new sophisticated statistical methods have to be developed and large prospective studies have to be undertaken. However, as the field of genetic epidemiology is growing exponentially, it is important to remember that cancer is foremost an environmental disease. We do not expect that one single genetic variant will have a useful predictive value in risk assessment. Many genetic determinants of cancer are most likely negligible compared to everyday exposures from the environment. In many cases, genetic variation will probably act as an effect modifier of certain exposures (e.g. UGT1A6 gene, aspirin use and colon adenoma) affecting the probability of developing disease rather than causing it. The significant improvement in public health will be by intervention through information about appropriate life style choices. With that said, the identification of low-penetrant genes is of utter importance. If we manage to understand why certain genes contribute to a specific
disease we might identify more directed treatments and effective prevention methods.

6.2 Future prospects – Genetic epidemiology of prostate cancer

Despite tremendous efforts during the last decade the genetic mechanisms involved in the pathogenesis and progression of prostate cancer remain to be identified. In the early 90’s, aggregation and segregation studies provided support for a hereditary component involved in disease aetiology. The first prostate cancer susceptibility locus (HPC1) was identified in 1996 through linkage analysis (88) and followed by identification of many suggestive loci such as PCAP, HPC2 and HPCX, all harbouring plausible genes. The success stories of BRCA1 and BRCA2 in breast cancer 1994-1995 led to a general opinion that it was only a matter of time until the responsible prostate cancer genes would be identified. When I started my doctoral training in 2004, the enthusiasm had been replaced by dejection and frustration. Linkage studies constantly reported novel susceptibility loci and there was a striking lack of replication. Association studies were only undertaken in low-powered studies leading to numerous spurious findings and failure of replication. At that time, there was an obvious lack of well-powered large studies but there has been a remarkable progress during the last year. BPC3 has launched their first publications and several case-controls studies with adequate power have been collected. BPC3 states in their research abstract in 2003 that they set out to investigate genetic polymorphisms in steroid hormone metabolizing genes and genes in the insulin-like growth factor pathway. So far, BPC3 has only reported results from two loci (HSD17B1 and 8q24) for prostate cancer but several papers are in the pipe-line. Collaborations between research groups such as BPC3 provide opportunities to perform well-powered studies with subsequent replication in independent populations and in different ethnicities. Large study populations are also essential for stratified analysis based on age and disease severity.

After years of unsuccessful efforts the first prostate cancer locus was identified a year ago (84). The identification of 8q24 was a result of a well conducted large-scale linkage analysis illustrating several important aspects: 1) Family-based studies are very effective in identifying genes as long as they are large-scale and well conducted. 2) Choosing a dense set of markers is crucial. There is a good chance of missing causal alleles if only a limited number of markers are assessed. 3) Our
knowledge about the biological mechanisms driving tumor development is limited. Candidate gene studies in prostate cancer have to date shown limited success. Chromosome 8q24 is a gene-poor region and researchers are at this moment searching for regulatory sequences and other functional variants that reside in this region.

As additional prostate cancer genes will be discovered we will obtain new tools to improve diagnostic and treatment strategies. So far, the vast majority of genetic association studies have considered risk rather than progression. It is my opinion that finding progression genes is of utter importance. Population-based case-control studies will in a few years constitute a solid base for performing survival analyses as long as well-documented follow-up is undertaken. Identifying genetic variants that contribute to rapid progression will ultimately help us to offer more effective treatments for those men who are in need of it.

6.3 Future implications based on this thesis

This thesis provides evidence of associations between genetic markers and prostate cancer risk and prognosis. To further explore these associations, replication in independent cohorts and studies in vivo as well as in vitro are required to fully understand the possible role of these genes. The pathway-driven approach that was undertaken in paper III shows a possible way of combining genes and thereby identify strong risk effects. However, as we only considered three genes in this pathway, expanded analysis including more genes is requested. The implication of AR in hormonal treatment response in paper IV should motivate other groups to pursue this hypothesis further and perform similar studies to confirm these findings. If confirmed it could ultimately result in more refined treatments.
6.4 Conclusions

Based on the studies included in this thesis, we can conclude the following:

- Common genetic variation in E-cadherin is important for prostate cancer development in families with multiple prostate cancer cases but not in the general population in Sweden.

- Earlier reported prostate cancer polymorphisms in AR, CYP17, GSTT1, MSR1 and SRD5A2 were replicated in the Swedish population.

- Common genetic variation in androgen regulating genes is associated with prostate cancer risk but not prognosis in the Swedish population.

- By combining main effects from several genetic variants in androgen regulating genes, population subgroups with a notable risk of developing prostate cancer can be identified.

- Common genetic variation in AR may affect hormonal treatment response and thereby prostate cancer-specific survival.

- Common genetic variation in the upstream region of the proto-oncogene ERG does not affect prostate cancer risk but may be important for prostate cancer-specific survival.
7. Populärvetenskaplig sammanfattning (Summary in Swedish)


En stor del av de individuella skillnaderna som finns mellan människor (som till exempel benägenheten att drabbas av sjukdom) kan förklaras av genetisk variation. Majoriteten av genetisk variation består av enstaka mutationer (single nucleotide polymorphisms, SNPar). Man beräknar att vår arvsmassa innehåller ungefär 12 miljoner SNPar, både inom och mellan de 25 000 gener vi bär på. Det finns starka bevis för att nedärvd genetisk variation har betydelse för både uppkomsten och prognosen av prostatacancer. Identifiering av dessa genetiska varianter skulle leda till ökad förståelse om de biologiska mekanismer som orsakar prostatacancer och därigenom bidra till utvecklandet av förebyggande åtgärder, diagnostisering samt behandling.

Syftet med denna avhandling är att studera hur genetisk variation i ett antal kandidatgener påverkar risken att utveckla och avlida av prostatacancer. Alla studier i avhandlingen bygger på en svensk populationsbaserad fall-kontrollstudie (CAPS). Vi samlade in blod från 2 965 svenska män som diagnostiserades med prostatacancer 2001-2003 samt 1 823 män utan prostatacancer. För att undersöka om genetisk variation i vissa specifika gener associerade med prostatacancerrisk och prognos jämförde vi förekomsten av olika SNPar hos sjuka och friska i fem olika delstudier.

I delstudie I undersökte vi om nedärvd genetisk variation i genen E-cadherin påverkade risken för prostatacancer. E-cadherin är en celladhesionsmolekyl som sitter på cellers yta och ser till att cellerna håller mycket tätt kontakt med
varandra. Slutsatsen från detta arbete är att vanlig genetisk variation i E-cadherin påverkar risken för prostatacancer i familjer med flera fall av prostatacancer men inte i den övriga befolkningen.

I delstudie II undersökte vi 46 genetiska varianter som tidigare har rapporterats påverka risken för prostatacancer. Vi lyckades replikera sex av de testade varianterna. En intressant observation var att tre av de associerade generna är inblandade i regleringen av manliga könshormoner.

Delstudie III var en fortsättning på delstudie II där vi undersökte de tre hormonrelaterade generna mer utförligt. Vi observerade att flera SNP:s associerade med prostatacancerrisken, speciellt för androgenreceptorn (ett signalprotein som svarar på signaler från manliga könshormoner). Män som bar på flera riskvarianter från de tre generna hade en kraftig riskökning.

I delstudie IV samlade vi in information om dödsorsaker för alla 2 965 prostatacancerfall i CAPS. När studien genomfördes hade sammanlagt 300 män dött på grund av prostatacancer. Ingen av de testade varianterna från delstudie III korrelerade med sjukdomsspecifik överlevnad. Vi utförde en delanalys på de män som hade fått hormonbehandling och observerade att de män som bar på en SNP i androgenreceptorn hade en ökad risk att dö av prostatacancer. Detta resultat indikerar att genetisk variation i androgenreceptorn kan påverka hur patienten svarar på hormonbehandling.

I delstudie V undersöktes om vanlig genetisk variation i genen ERG påverkade risken eller prognosen för prostatacancer. ERG är en så kallad onkogen som vid mutationer ger upphov till tumörceller. Vi observerade ingen skillnad i genetisk variation när vi jämförde sjuka med kontrollgruppen. Däremot såg vi ett samband mellan genetisk variation 100 000 baspar bort från ERG och prognos.

Sammanfattningsvis bidrar denna avhandling till ökad förståelse om hur vanlig genetisk variation påverkar risken att drabbas av avlida i prostatacancer i Sverige. Vi såg indikationer på att genetisk variation i hormonrelaterade gener påverkar risken för att utveckla prostatacancer. Dessutom kan genetisk variation i en av dessa gener påverka hur bra patienten svarar på hormonbehandling vid avancerad prostatacancer. Ytterligare forskning krävs för att verifiera dessa resultat.
8. Acknowledgements

Many people have contributed to this thesis. I would like to express my sincerest
gratitude to you all. In particular I would like to thank:

All study participants. Without your generous contribution this research
would never been possible from the beginning.

Henrik Grönberg, my main supervisor. Thank you for introducing me
to science and giving me the opportunity to work in the intriguing field of genetic
epidemiology. Your broad knowledge in cancer genetics has taught me more than I
thought was possible. Your positive attitude, dedication to research and belief in me
has been a never-ending source of inspiration. It has been a true privilege.

Fredrik Wiklund, my co-supervisor for your never failing support and
encouragement. You have always been there for all questions, big and small. Without
your statistical and methodological expertise I would have been at a loss.

Mattias Johansson, my dear friend and colleague. No one knows me
better than you and still you are always by my side. Thank you for friendship,
support and for being the one I have shared my everyday work with these years. You
are the best travel partner ever! Who is going to keep track of me now?

The staff at Onkologiskt Centrum for making it fun to go to work.
Thanks to you the coffee breaks get a little bit longer than anticipated. Lena for
welcoming me to OC and providing me an office space during these years. Åsa for
all help with final preparation of this thesis, for all music we shared and for being a
friend. Benjamin for daring to share office with me these last weeks.

Everyone in the research group. Bettan and Karin for your tireless effort
in identifying family members. Björn-Anders and Monica for always answering my
questions about genetics. Lena for help with everything from faxes to applications.
Bea for your positive attitude and for letting me be a part of your projects. Camilla,
Fredrik, Katarina, Kristina and Tanja for your enthusiasm in my projects and for
struggling with your own showing that it is doable.

All co-authors, collaborators and CAPS-people for your contribution to
this work. Hans-Olov Adami for your valuable comments. The staff at the
Department of Medical Epidemiology and Biostatistics at Karolinska Institutet for
creating a truly inspiring research environment.
Carina and Anna for all administrative help. You know everything that is worth knowing.

All inspiring people I have met around the world during courses and conferences. Kristina for your kindness and support regarding my future academic career.

All my physics friends for the university years we spent together. You are all too far away. I miss you! Ia for always being only a phone-call away.

All my friends in Umeå for friendship, sporting events, game and movie nights, dinners, barbecues, parties... Without you, these years would have been much less fun!

Anna, Hanna, Lena and Maria for being there no matter what. Thank you for loving me as I am and for giving me energy when I need it the most (and all other times as well). You are the best friends anyone could ever wish for! Anna for all (talkative) squash games and for sharing my interest in research. Hanna for your infectious untroubled attitude and for being the one I can call 24/7. Lena for dragging me outside in the winter (although I am reluctant) and for always making me smile. Maria for your ability to make everything self-evident and for always making me feel that things are going to be ok.

My brother Jonas for still being my idol. Our discussions always make me stop and reflect; a rare event in my life. You and Linda have always kept your door generously open when I have been working in Stockholm. Hugo, for letting me experience the world through a child’s eye again.

My parents Ulf and Britta for unconditional love and support in all aspects of life. Thank you for giving me and my brother a happy and caring childhood and for providing us with everything we have ever needed in life.

Jessica. Thank you for listening to me and encouraging me. Thank you for teaching me medicine and for all fruitful discussions about everything in life including research. And most of all, thank you for loving me and always reminding me that life is more than work. I love you!

This work has been supported by NIH, the Swedish Cancer Society (Cancerfonden) and Umeå University.
9. References


54. The National Board of Health and Welfare, Sweden. URL: http://www.socialstyrelsen.se


98. Severi, G, Hayes, VM, Padilla, EJ, et al. The common variant rs1447295 on chromosome 8q24 and prostate cancer risk: results from an Australian


156. StataCorp Stata Statistical Software. Release 8 College Station, TX: StataCorp LP; 2005.


