RIIKKA NURMINEN
Genetic Susceptibility Factors for Prostate Cancer at Chromosomal Region 11q13.5
RIIKKA NURMINEN

Genetic Susceptibility Factors for Prostate Cancer at Chromosomal Region 11q13.5

ACADEMIC DISSERTATION
To be presented, with the permission of the Board of the BioMediTech of the University of Tampere, for public discussion in the auditorium of Finn-Medi 5, Biokatu 12, Tampere, on 10 June 2016, at 12 o’clock.

UNIVERSITY OF TAMPERE
RIIKKA NURMINEN

Genetic Susceptibility Factors for Prostate Cancer at Chromosomal Region 11q13.5

Acta Universitatis Tamperensis 2170
Tampere University Press
Tampere 2016
SUPPLEMENTARY MATERIALS

Supervised by
Professor Johanna Schleutker
University of Turku
Finland
Docent Tiina Wahlfors
University of Tampere
Finland

Reviewed by
Docent Peter Boström
University of Turku
Finland
Docent Pia Vahteristo
University of Helsinki
Finland

The originality of this thesis has been checked using the Turnitin OriginalityCheck service in accordance with the quality management system of the University of Tampere.

Copyright ©2016 Tampere University Press and the author

Cover design by
Mikko Reinikka

Distributor:
verkkokauppa@juvenesprint.fi
https://verkkokauppa.juvenes.fi

Acta Universitatis Tamperensis 2170
ISBN 978-952-03-0125-5 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1669
ISSN 1456-954X
http://tampub.uta.fi

Suomen Yliopistopaino Oy – Juvenes Print
Tampere 2016
Contents

List of Original Communications ......................................................................................... 5
Abbreviations ...................................................................................................................... 6
Abstract ............................................................................................................................... 9
Tiivistelmä .......................................................................................................................... 11

1  Introduction .................................................................................................................... 13

2  Review of the Literature ............................................................................................... 15
  2.1  Prostate cancer .......................................................................................................... 15
        2.1.1  Incidence and mortality ...................................................................................... 15
        2.1.2  Risk factors ........................................................................................................ 16
        2.1.3  Prostate cancer stage ........................................................................................ 18
        2.1.4  Biomarkers ......................................................................................................... 19
  2.2  Genetic susceptibility of prostate cancer .................................................................... 23
        2.2.1  Heritability ......................................................................................................... 23
        2.2.2  From linkage analyses to post-GWAS approaches ........................................... 26
        2.2.3  11q13-14 .......................................................................................................... 32
  2.3  EMSY .......................................................................................................................... 37

3  Aims of the Study ........................................................................................................ 40

4  Materials and Methods .............................................................................................. 41
  4.1  Study subjects (I, II and III) ................................................................................... 41
        4.1.1  Samples ............................................................................................................. 41
        4.1.2  Classifying study subjects for association testing ........................................... 42
  4.2  Laboratory methods (I, II and III) ........................................................................... 44
  4.3  Computational methods ............................................................................................ 46
        4.3.1  Tag SNP determination, imputation and association testing
              of imputed variants (II) ......................................................................................... 46
        4.3.2  Association testing of genotyped variants (I and II) ........................................ 46
        4.3.3  Haplotype analysis and LD (I, II and III) ......................................................... 47
        4.3.4  eQTL analysis (III) ............................................................................................ 47
        4.3.5  Functional annotation ....................................................................................... 48
List of Original Communications

This thesis is based on the following communications, referenced in the text by their Roman numerals (I-III).


The original publications have been reproduced with the permission of the copyright holders.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACER3</td>
<td>alkaline ceramidase 3</td>
</tr>
<tr>
<td>AKT1</td>
<td>v-akt murine thymoma viral oncogene homolog 1</td>
</tr>
<tr>
<td>AR</td>
<td>androgen receptor</td>
</tr>
<tr>
<td>ARL11</td>
<td>ADP-ribosylation factor like GTPase 11</td>
</tr>
<tr>
<td>AXIN2</td>
<td>axin 2</td>
</tr>
<tr>
<td>BRCA2</td>
<td>breast cancer 2</td>
</tr>
<tr>
<td>CBX1</td>
<td>chromobox 1</td>
</tr>
<tr>
<td>CCND1</td>
<td>cyclin D1</td>
</tr>
<tr>
<td>CEBPB</td>
<td>CCAAT/enhancer binding protein beta</td>
</tr>
<tr>
<td>CHEK2</td>
<td>checkpoint kinase 2</td>
</tr>
<tr>
<td>CLIA</td>
<td>Clinical Laboratory Improvement Amendment</td>
</tr>
<tr>
<td>CRPC</td>
<td>castration-resistant prostate cancer</td>
</tr>
<tr>
<td>dbSNP</td>
<td>the Single Nucleotide Polymorphism Database</td>
</tr>
<tr>
<td>DGAT2</td>
<td>diacylglycerol O-acyltransferase 2</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>ELAC2</td>
<td>elaC ribonuclease Z 2</td>
</tr>
<tr>
<td>EMSY</td>
<td>EMSY, BRCA2-interacting transcriptional repressor</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopedia of DNA Elements</td>
</tr>
<tr>
<td>ENT</td>
<td>EMSY N-terminal</td>
</tr>
<tr>
<td>eQTL</td>
<td>expression quantitative trait locus</td>
</tr>
<tr>
<td>ERSPC</td>
<td>the European Randomized Study of Screening for Prostate Cancer</td>
</tr>
<tr>
<td>ETS1</td>
<td>v-ets avian erythroblastosis virus E26 oncogene homolog 1</td>
</tr>
<tr>
<td>FAM57A</td>
<td>family with sequence similarity 57 member A</td>
</tr>
<tr>
<td>FDA</td>
<td>the U.S. Food and Drug Administration</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>FOXA1</td>
<td>forkhead box A1</td>
</tr>
<tr>
<td>GTEx</td>
<td>Genotype-Tissue Expression</td>
</tr>
<tr>
<td>GWAS</td>
<td>genome-wide association study</td>
</tr>
<tr>
<td>HapMap</td>
<td>haplotype map</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------------------</td>
</tr>
<tr>
<td>SNP</td>
<td>single-nucleotide polymorphism</td>
</tr>
<tr>
<td>SR</td>
<td>serine/arginine</td>
</tr>
<tr>
<td>STHLM3</td>
<td>Stockholm 3</td>
</tr>
<tr>
<td>tag SNP</td>
<td>tagging SNP</td>
</tr>
<tr>
<td>TCF</td>
<td>T-cell factor</td>
</tr>
<tr>
<td>TURP</td>
<td>transurethral resection of the prostate</td>
</tr>
<tr>
<td>UBP1</td>
<td>upstream binding protein 1 (LBP-1a)</td>
</tr>
<tr>
<td>VPS53</td>
<td>vacuolar protein sorting 53 homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>ZMYND11</td>
<td>zinc finger MYND-type containing 11</td>
</tr>
</tbody>
</table>
Prostate cancer is a significant public health concern worldwide. It is the most common cancer affecting men in Finland and in other Western countries. The increased incidence of prostate cancer can be at least partly explained by the development of prostate-specific antigen testing, which results in a high rate of overdiagnosis and overtreatment. Therefore, improved biomarkers are needed to identify men at high risk of aggressive prostate cancer among the majority of patients for tailored monitoring and treatment of this disease.

The main risk factors for prostate cancer are increased age, family history and ethnicity. Inherited genetic factors contribute to the risk of prostate cancer, which has the highest estimated genetic contribution among common cancers. Studies of prostate cancer susceptibility have suggested that prostate cancer is a genetically complex disease with susceptibility factors located in multiple chromosomal regions. The chromosomal region 11q13-14 has been found to contain prostate cancer-predisposing factors in both linkage and association studies. This region has been specifically linked to prostate cancer in Finnish prostate cancer families, and it has been reported to contribute to the aggressive form of this disease. No candidate gene or other causal factor has been identified in the region to date. One interesting gene in this region is EMSY, located at 11q13.5, which has been identified as a candidate gene for breast and ovarian cancers.

The aim of the thesis was to study the region 11q13.5 in relation to prostate cancer susceptibility in Finnish men. Samples and patient information from Finnish prostate cancer patients and samples from male controls were obtained for use in this study. Fine mapping of this region using imputation and more precise screening of EMSY by Sanger sequencing resulted in identification of three intronic EMSY single-nucleotide polymorphisms (SNPs) and six intergenic variants that predispose to prostate cancer. A rare EMSY mutation was found to increase the risk, particularly that of aggressive prostate cancer, in Finnish population, and it was detected in men in Finnish prostate cancer families with aggressive disease. Intergenic common variants, which were correlated with each other, were strongly associated with the risk of prostate cancer death. In addition, haplotypes including the identified risk SNPs were found to contribute to disease predisposition.
Functionality of the variants was assessed by functional annotation and by examining effects of the variants on gene expression by expression quantitative trait loci (eQTL) analysis. The prostate cancer death-associated risk SNPs coincided with enhancer elements in multiple cell types and were observed to affect the expression of \textit{DGAT2} in prostate tumours and of \textit{AP001189.4} in whole blood, suggesting tissue-specific gene regulation as a mechanism promoting tumour development. The functional annotations indicated that the \textit{EMSY} SNPs may affect messenger RNA splicing of \textit{EMSY}, but this finding warrants experimental confirmation.

In conclusion, alterations in the chromosomal region 11q13.5 contribute to prostate cancer susceptibility in Finnish men, increasing the risk particularly of aggressive cancer and life-threatening disease progression. Replication of the associations of these variants in other populations, as well as additional analyses of the identified target genes, are necessary for detailed characterization of their tumour-promoting properties. This thesis has revealed novel information regarding prostate cancer susceptibility that could be used in the development of a biomarker panel specific for detection of an increased risk of aggressive and advanced prostate cancer.
Eturauhassyövän kansanterveydellinen merkitys on suuri. Se on yleisin miesten syöpä niin Suomessa kuin myös muissa länsimaisissa. Syöpätapausten määrä kasvaminen selittyy osittain eturauhassepsifisen antigenin käyttämisellä biologisena merkkiaineena eli biomarkkerina syövän diagnostiikassa, koska kyseisellä menetelmällä havaitaan myös hoitoa tarvitsemmattomia syöpiä. Uusia biomarkkeriteita tarvitaan tunnistamaan eturauhassyöpään sairastuvista miehistä ne, joilla on kohonnut riski sairastua nopeasti etenevään aggressiiviseen syöpätyyppiin, mikä mahdollistaisi tehokkaamman yksilöidyn hoitosuunnitelman tekemisen.


Väitöskirjatyössä selvitettiin kromosomialueen 11q13.5 osuutta eturauhassyöpälaitteeseen suomalaisilla. Tutkimuksessa käytettiin aineistona näytteitä ja potilastietoja suomalaisilta eturauhassyöpäpotilaisilta ja mieskontrollilleilta. Aluetta tutkittiin etsimällä syövälle altistavia muutoksia imputointi-menetelmällä ja tunnistamalla EMSY-geenin muutoksia sekvensoimalla. Tutkimuksessa havaittiin kolme intronista muutosta EMSY-geenissä ja kuusi geenien välisellä alueella sijaitseva variantti, jotka yhdistettiin kohonneeseen syöpälaitteeseen. Harvinainen, introninen EMSY-muutos altisti etenkin aggressiiviselle syövälle suomalaisessa populaatiossa, ja sen havaittiin esiintyvän eturauhassyöpäperheissä miehillä, joilla on aggressiivinen tauti, kun taas frekvenssiltään yleiset, geenien välisellä alueella sijaitsevat keskenään korreloituneet variantit yhdistettiin eturauhassyöpäkuoleman...
riskiin. Yksittäisten geneettisten muutosten lisäksi alueella tunnistettiin syöpäalttiuteen liittyviä haplotyyppejä.


1 Introduction

Cancer is a major health problem worldwide; millions of new cancer cases are diagnosed annually and tens of millions of people are living with cancer (Bray et al., 2013; Ferlay et al., 2013a). It is the second leading cause of death in the US and is anticipated to soon surpass heart disease as the top-ranking cause (Siegel et al., 2015). Further, it is a complex disease caused by multiple environmental, lifestyle and inherited genetic factors.

Cancer results from uncontrolled growth and proliferation of cells (Hanahan and Weinberg, 2000). Although multiple types of cancer exist, cancer cells share similar malignant growth-promoting properties, such as the ability to evade programmed cell death (apoptosis), self-sufficiency of growth signals, insensitivity to anti-growth signals, limitless replicative potential, sustained angiogenesis, the capacity to invade adjacent tissues and to metastasize, the capacity for reprogramming energy metabolism and the ability to evade immune destruction (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011). The accumulation of mutations and chromosomal rearrangements enable cancer cells to acquire the properties that drive tumourigenesis (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011). In addition, inflammation is thought to contribute to the tumourigenic properties of cells (Hanahan and Weinberg, 2011). Mutations accumulate throughout life, and thus, the onset of many cancers is age-dependent (Hanahan and Weinberg, 2000). Both acquired and inherited genetic alterations result in the dysfunction of genes, which are commonly divided into two main classes in cancer biology: oncogenes and tumour suppressor genes. Oncogenes, which are over-activated in cancer cells, promote cell proliferation (Todd and Wong, 1999), and tumour suppressor genes, which are inactivated in cancer cells, regulate cell cycle progression (Kinzler and Vogelstein, 1997). Accumulation of five to seven rate-limiting events, i.e., driver mutations that confer selective growth advantage, is considered to be required for a tumour initiation and progression (Stratton et al., 2009); however, it has been recently suggested that only three driver mutations would be sufficient for the development of lethal cancer (Tomasetti et al., 2015).

Genetic susceptibility to cancer is the result of inheritance of genetic factors that increase the risk of cancer development, because fewer mutations are required for a
cell to develop the capacity for malignant growth when inherited alterations are present (Vogelstein and Kinzler, 2004). However, the presence of an increased risk does not necessarily mean that cancer will develop, because the inheritance of genetic factors is not sufficient to directly cause cancer; additional somatic mutations are required for cancer to develop (Vogelstein and Kinzler, 2004). The contribution of heritable factors to the development of many common cancers is minor, but a strong genetic component has been described for prostate, colorectal and breast cancers, among which prostate cancer is the most heritable (Lichtenstein et al., 2000). Studies of genetic susceptibility aim to identify risk factors that can be utilized for cancer prevention and risk assessment, as well as to increase the knowledge of cancer development, which can be applied in the treatment of cancer patients.
2  Review of the Literature

2.1  Prostate cancer

The prostate is an organ of the male reproductive system that is situated caudal to the bladder and surrounds the proximal urethra. It produces and secretes seminal fluid containing prostate-specific antigen (PSA), which modifies the structure of semen so that it is more fluid (Lilja et al., 1987). Prostate tumours typically originate from epithelial cells of this gland and are thus called adenocarcinomas (Bracarda et al., 2005). The development of prostate cancer is commonly multifocal, meaning that several tumour foci exist simultaneously (Andreoiu and Cheng, 2010).

2.1.1  Incidence and mortality

Prostate cancer is the second most common cancer affecting men worldwide (Ferlay et al., 2013a). It was estimated to account for approximately 26% (n = 220,800) of new cancer diagnoses in 2015 in men in the US making it the most frequently diagnosed male cancer in this country (Siegel et al., 2015). Similarly, 5124 new cases were diagnosed in Finnish men in 2013, resulting in highest rate of diagnosis (31%) among all primary cancers in Finland (Finnish Cancer Registry). The incidence of prostate cancer in Finland increased steadily from the 1950s to the beginning of the 1990s (Finnish Cancer Registry) (Figure 1). During the past 20 years this incidence has doubled due to the development of PSA testing as a screening method in the 1990s, improvements in imaging and biopsy technologies and lifestyle factors such as westernization of diet and sedentary behaviour (Center et al., 2012; Finnish Cancer Registry) (Figure 1). The incidence is expected to increase in Finland due to the long average life expectancy of men and because men born during the post-world war II baby boom are reaching the average age of onset of this disease (Pukkala et al., 2011).

Prostate cancer is the second leading cause of cancer death in men, accounting for 12% of all cancer deaths in Finnish men in 2013 (Finnish Cancer Registry), and it was estimated to account for 9% of all cancer deaths in the US in 2015 (Siegel et al., 2015). The mortality rate of this disease has not substantially changed over the
years (Finnish Cancer Registry) (Figure 1). The 1-year and 5-year survival rates of prostate cancer in Finland are 98% and 93%, respectively, which are among the highest rates reported for all types of cancer (Finnish Cancer Registry).

![Figure 1. Incidence and mortality rates of prostate cancer in Finland from 1967-2013. The data were modified from the Finnish Cancer Registry.](image)

2.1.2 Risk factors

Age. The occurrence of prostate cancer is extremely age-dependent. Its incidence is increased in men over 40-years of age (Finnish Cancer Registry), and the average age at diagnosis in Finnish men is approximately 70 years (Pukkala et al., 2011). The probability of developing prostate cancer is the highest in men over 70 years (11%) (Siegel et al., 2015).

Family history. Prostate cancer cases have been observed to cluster in families. This clustering can result from inherited susceptibility, lifestyle habits, and other environmental factors and their interactions (Verhage and Kiemenev, 2003). The risk of prostate cancer increases with an increasing number of affected first-degree relatives, with increases of 2- to 3-fold in men with a single affected relative and of over 4-fold in men with two or more affected family members (Kicinski et al., 2011). In addition, the risk is higher in men with an affected brother compared to those with an affected father (Kicinski et al., 2011). Furthermore, family history
predisposes individuals to early-onset of this disease (before age 65) (Kicinski et al., 2011). In a Finnish prostate cancer family-based study, the risk was determined to be increased by approximately 2- to 3-fold in the first-degree relatives of patients with either early (before age 70) or late disease onset (after age 80) (Matikainen et al., 2001). A specific late-onset predisposing genetic factor could explain these findings (Matikainen et al., 2001).

Prostate cancer cases are divided into three categories based on family history. The majority of cases (75-85%) are sporadic, with no family history of disease (Ostrander et al., 2004). Both familial and hereditary prostate cancer cases have a positive family history, but the criteria for identifying hereditary cases are stricter with regard to cancer occurrence in a pedigree (Carter et al., 1993). Familial cases account for approximately 10-20% of all prostate cancer cases, and hereditary cases account for approximately 5-10% (Carter et al., 1993). However, despite the definitions, sporadic prostate cancer also has a germline genetic component (Lichtenstein et al., 2000; Lu et al., 2014).

**Geographic differences and ethnicity.** Most prostate cancer cases (70%) are diagnosed in more developed regions (Ferlay et al., 2013a). Australia, New Zealand, Northern America and Western and Northern Europe have the highest prostate cancer incidences (Ferlay et al., 2013a). Most regions in Asia and Northern Africa have 10-20-fold lower incidences compared to the high-rate regions (Ferlay et al., 2013a). Prostate cancer is diagnosed more frequently in Nordic countries, including Finland, compared to most European countries (Ferlay et al., 2013b). In the US, the incidence varies among ethnic groups, with the highest rates observed in non-Hispanic black men, followed by non-Hispanic white and Hispanic men, who have more than 1.5-fold lower incidences compared to black men (Siegel et al., 2015). Further, American Indian/Alaska Native and Asian/Pacific Islander men have been reported to have approximately 2- to 3-fold lower incidences than non-Hispanic black men (Siegel et al., 2015). Various factors have been suggested to explain the observed regional and ethnic variabilities in prostate cancer risk such as environmental and lifestyle factors, the availability of health-care, screening patterns and genetic factors (Hsing et al., 2000).

**Other factors.** Age, family history and ethnicity are the major prostate cancer risk factors, but other factors have been reported to have potential significance in cancer prevention. With regard to dietary factors, high intake of saturated fat, well-done meat and calcium may increase the risk of advanced prostate cancer, while the influences of total meat, fruit and vegetable intake on this risk are unclear (Gathirua-Mwangi and Zhang, 2014). In addition, studies on the levels of hormones,
particularly that of testosterone, have reported inconsistent results (Klap et al., 2015). Accumulating data suggest that lifestyle factors, such as obesity (Allott et al., 2013) and smoking (Islami et al., 2014), may be associated with aggressive prostate cancer and prostate cancer death, and alcohol consumption has also been correlated with an overall increased risk (Bagnardi et al., 2015); however, further studies are warranted to support these findings.

Chronic inflammation has been suggested to contribute to the risk of developing prostate cancer (De Marzo et al., 2007). Prostatic inflammation-promoting factors are largely unknown, but potential causes of prostate inflammation include infectious agents, urine reflux, physical trauma, hormonal changes and dietary habits [reviewed in (De Marzo et al., 2007)]. The average prevalence of symptoms of prostatitis syndromes, including bacterial infections, both inflammatory and non-inflammatory chronic pelvic pain syndromes and asymptomatic inflammation, in men is 8% (Krieger et al., 2008); however, the overall rate is anticipated to be much higher due to asymptomatic conditions (Jiang et al., 2013). Few studies have directly examined the association between chronic inflammation and prostate cancer risk, but inflammation of benign prostate tissues has been suggested to predispose individuals, especially to aggressive prostate cancer (Gurel et al., 2014). In addition, a significant association has been observed between self-reported prostatitis and prostate cancer (Jiang et al., 2013). Interestingly, multiple inherited genetic variants related to inflammatory pathways have been found to increase the risk of prostate cancer, but the effects of these variants on intraprostatic inflammation as a prostate cancer risk-increasing mechanism warrant further study [reviewed in (De Marzo et al., 2007)].

### 2.1.3 Prostate cancer stage

The probability of developing prostate cancer during a lifetime is 15% (1/7) (Siegel et al., 2015). Prostate cancer exists as both a symptomatic and asymptomatic, i.e., latent, disease. The average asymptomatic period before the appearance of symptoms and clinical diagnosis of prostate cancer has been estimated to be 11-12 years (Etzioni et al., 1998). Prior to PSA testing, the proportion of latent prostate cancer detected at autopsy was approximately 20-40% (Breslow et al., 1977; Holund, 1980; Yatani et al., 1982). The occurrence of latent prostate cancer detected at autopsy has decreased by 3-fold, and the proportion of lower-grade latent cancer has increased due to PSA screening (Konety et al., 2005). Approximately 81% of prostate
tumours are diagnosed as localized, 12% as regional and 4% as distal (Siegel et al., 2015). Localized and regional cancers have good prognoses, with relative five-year survival rates of > 99%, whereas the five-year survival rate of distal, metastasized cancer is only 28% (Siegel et al., 2015).

The aggressiveness of prostate tumours is commonly described using histologic Gleason grading, with grades ranging from 2-10 (Epstein, 2010; Gleason and Mellinger, 1974). The grade is determined as the sum of the primary and secondary patterns observed in a specimen. Higher grades represent less differentiated tumours and indicate worse prognosis. No consensus for defining aggressive prostate cancer based on Gleason grading exists; both Gleason grades of equal to and those of greater than 7 and 8 are used (Sartori and Chan, 2014).

2.1.4 Biomarkers

Biological markers, i.e., biomarkers, are measurable variables and include proteins, metabolites, ribonucleic acid (RNA) and deoxyribonucleic acid (DNA), and variations of them and in their levels are indicative of the medical state of a patient (Prensner et al., 2012). Biomarkers can be used to determine the risk or prognosis of a disease, to screen for and diagnose a patient with a disease or to choose and monitor a treatment (Prensner et al., 2012).

**PSA.** PSA is the most commonly used biomarker for prostate cancer. PSA screening was first performed in the 1980s, when it replaced prostatic acid phosphatase measurement (Ercole et al., 1987; Prensner et al., 2012). PSA is a prostate-specific serine protease (Watt et al., 1986) that cleaves proteins in seminal fluid and affects the structure of semen (Lilja et al., 1987). In prostate cancer patients, the prostate tissue structure is damaged due to a tumour formation, which has been suggested to result in an increase in the PSA level in the blood; however, the precise underlying mechanism remains unclear (Lilja et al., 2008).

A threshold level of 4 ng/mL PSA has been traditionally considered to warrant further examination of a patient (Catalona et al., 1994). Although a higher PSA level indicates a higher probability of developing prostate cancer (Heidenreich et al., 2014), a low level (< 4 ng/mL) is detected in 12.5-25% of individuals diagnosed with high-grade prostate cancer (Thompson et al., 2004). Factors other than cancer, such as benign prostatic hyperplasia, prostatitis, age, body mass index and race, affect the serum PSA level (Lilja et al., 2008), indicating that PSA is not a prostate cancer-specific molecule. Modified PSA measurements, such as measurements of PSA
density, age-specific levels and the ratio of different forms of PSA, have been suggested for improved PSA testing (Lilja et al., 2008). In addition to screening, PSA measurements can be used at and after prostate cancer diagnosis. Prostate cancer patients can be divided into several risk categories based on the serum PSA level, in addition to the clinical stage and biopsy Gleason grade, to assess the treatment options and prognosis of localized prostate cancer (Heidenreich et al., 2014; Prostate cancer: Current Care Guidelines Abstract, 2014). Furthermore, PSA can be used to monitor patients for cancer recurrence. Further examination and treatment options are assessed if an elevated PSA level, i.e., biochemical recurrence, is detected after a prostate cancer treatment (Prostate cancer: Current Care Guidelines Abstract, 2014).

Prostate cancer screening is not performed nationwide in Finland, in contrast with breast, cervical and colorectal cancer screening (Finnish Cancer Registry). However, two major randomized controlled screening trials of prostate cancer have reported findings regarding PSA-based screening: the European Randomized Study of Screening for Prostate Cancer (ERSPC) and the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The ERSPC recruited 182,000 men from seven European countries, including the Netherlands, Sweden, Finland, Belgium, Spain, Italy and Switzerland (Schroder et al., 2009), and the PLCO was a US population-based trial that included 76,000 men (Andriole et al., 2009). Both trials were launched to determine whether PSA-based screening results in reduced cancer mortality (Andriole et al., 2009; Schroder et al., 2009). In the PLCO trial, which included 13 years of follow-up, no evidence of a reduction in prostate cancer mortality due to organized annual screening was detected (Andriole et al., 2012), whereas in the ERSPC trial, a substantial reduction of 21% was observed (Schroder et al., 2014). However, to prevent one death, a total of 781 men would need to be screened, and 27 prostate cancer cases would have to be diagnosed (Schroder et al., 2014). The differences in results between these two trials are thought to be due to contamination of the control arm with PSA screening in the PLCO trial (Pinsky et al., 2010). The main negative consequence of screening is overdiagnosis, i.e., diagnosis of a man with latent cancer that would not have been detected during his lifetime without the screening (Schroder et al., 2009). The age-dependent overdiagnosis rates of screening-detected cancer have been estimated to range from 27 to 56% for a single screening test and to be 48% and 50% for screening performed at four-year intervals and annually, respectively (Draisma et al., 2003). Overall, the estimates of prostate cancer overdiagnosis have varied from 1.7% to 67%, depending on the study design (Loeb et al., 2014). Overdiagnosis may lead to unnecessary treatment, which may cause patients to experience physical and mental adverse
effects to in addition to major costs to the healthcare. Targeted screening of men with a higher genetic risk of prostate cancer has been suggested for reducing the rate of overdiagnosis (Pashayan et al., 2015a; Pashayan et al., 2015b).

**Genomic and proteomic-based clinical tests for prostate cancer.** The findings of genomic and proteomic research have been used in the development of commercial tests for prostate cancer to improve screening and to guide treatment decisions (Table 1). Most of the U.S. Food and Drug Administration (FDA)-approved tests, such as the Prostate Health Index (phi) (Table 1), involve measurement of different forms of the PSA protein (U.S. Food and Drug Administration). The analytical and clinical performances of the FDA-approved tests have been reviewed to establish their acceptance as diagnostic tests (Sartori and Chan, 2014). In addition to FDA-approved tests, other, less validated tests are available and they are offered under a laboratory’s Clinical Laboratory Improvement Amendment (CLIA) certificate (Saini, 2016; Sartori and Chan, 2014) (Table 1). Some of the recently developed tests have shown to be useful in determining the risk of aggressive cancer (Table 1), and the currently on-going research is anticipated to produce findings that may be used for improved prediction of aggressive prostate cancer, enabling its early detection and proper treatment (Sartori and Chan, 2014).
Table 1. Commercial genomic and proteomic-based tests for prostate cancer (data obtained from Saini 2016).

<table>
<thead>
<tr>
<th>Testa</th>
<th>Company</th>
<th>Sample</th>
<th>Markers</th>
<th>Predicts aggressive cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before diagnosis, who needs biopsy after PSA testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate Health Index ( \textit{phi} )b</td>
<td>Beckman Coulter</td>
<td>blood</td>
<td>variants of PSA</td>
<td>yesc</td>
</tr>
<tr>
<td>4Kscore Test</td>
<td>Opko Health Inc.</td>
<td>blood</td>
<td>hK2 and variants of PSA</td>
<td>yes</td>
</tr>
<tr>
<td>Mi-Prostate Score</td>
<td>University of Michigan MLabs</td>
<td>blood/urine</td>
<td>PSA, PCA3 and TMPRSS2-ERG RNAs</td>
<td>yes</td>
</tr>
<tr>
<td>Before diagnosis, who needs rebiopsy after negative biopsy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progena PCA3 Assayb</td>
<td>Hologic</td>
<td>urine</td>
<td>PCA3 and PSA RNAs</td>
<td>no</td>
</tr>
<tr>
<td>ConfirmMDx</td>
<td>MDx Health Inc.</td>
<td>biopsy</td>
<td>methylation of GSTP1, APC and RASSF1</td>
<td>no</td>
</tr>
<tr>
<td>Prostate Core Mitomic Test</td>
<td>Mitomics</td>
<td>biopsy</td>
<td>mtDNA deletions</td>
<td>no</td>
</tr>
<tr>
<td>At diagnosis, to guide treatment decisions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oncotype DX Genomic Prostate Score</td>
<td>Genomic Health Inc.</td>
<td>biopsy</td>
<td>RNA of 17 genes</td>
<td>yes</td>
</tr>
<tr>
<td>ProMark</td>
<td>Metamark</td>
<td>biopsy</td>
<td>eight proteins</td>
<td>yes</td>
</tr>
<tr>
<td>Prolaris</td>
<td>Myriad Genetics Inc.</td>
<td>biopsy</td>
<td>RNA of 46 genes</td>
<td>yes</td>
</tr>
<tr>
<td>After surgery, to guide treatment decisions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decipher</td>
<td>Genome Dx Biosciences</td>
<td>tissue</td>
<td>RNA of 22 genes</td>
<td>yesd</td>
</tr>
</tbody>
</table>

a The test is offered under a laboratory’s CLIA certificate unless otherwise specified.
b FDA-approved test
c The test may be useful in determining aggressive prostate cancer risk but further validation is needed (Wang et al., 2014).
d The test predicts the probability of metastatic disease after radical prostatectomy.

Abbreviations: mtDNA = mitochondrial DNA, PSA = prostate-specific antigen
2.2 Genetic susceptibility of prostate cancer

2.2.1 Heritability

Heritability is an estimate of the proportion of phenotypic variation of a particular trait in a population explained by inherited genetic variations among individuals (Lichtenstein et al., 2000); the remainder of phenotypic variation is due to environmental factors. The contribution of genetic components to a disease can be investigated using twin-based studies, in which concordance rates for the disease are compared between monozygotic (identical) and dizygotic (non-identical) twins. Estimates of the heritability of prostate cancer have ranged from 16% to 45% (Baker et al., 2005; Lichtenstein et al., 2000), and the most recent estimate was 58% (Hjelmborg et al., 2014). Prostate cancer is considered the most heritable type of common cancer (Lichtenstein et al., 2000).

Segregation analysis is used to determine the mode of inheritance of a trait using family data. Multiple inheritance patterns have been suggested for prostate cancer, including autosomal dominant (Carter et al., 1992; Cui et al., 2001; Gronberg et al., 1997; Schaid et al., 1998), autosomal recessive and X-linked patterns (Cui et al., 2001; Monroe et al., 1995), and various estimates of penetrance and frequency of genetic variants have been reported. Penetrance refers to the extent to which a genetic variant affects expression of a disease phenotype, i.e., the proportion of all mutation carriers who express the disease phenotype. The autosomal dominant mode of inheritance has been associated with early-onset prostate cancer (Carter et al., 1992; Cui et al., 2001; Schaid et al., 1998), and the recessive and X-linked inheritance patterns have been correlated with an older age at diagnosis (Cui et al., 2001). Finnish population-based analysis has indicated that recessive inheritance associate with both early and late disease onset and that a polygenic or multifactorial component contribute to the prostate cancer risk rather than a single mutation (Pakkanen et al., 2007). Similarly, other reports have described polygenic models consisting of multiple recessively inherited genes with strong effects and of several genes with smaller multiplicative effects (MacInnis et al., 2010) or consisting mainly of multiple low-penetrance genes (Gong et al., 2002). In addition, differences in the penetrance of susceptibility factors have been identified among ethnicities (Gong et al., 2002).
Indeed, an increasing number of genetic studies have suggested that genetic predisposition to prostate cancer is very complex. According to the Common Disease, Common Variant hypothesis, the genetic susceptibility of a high-incidence disease, such as prostate cancer, can be explained by high-frequency variants with low penetrance (Reich and Lander, 2001; Schork et al., 2009). On the other hand, the Common Disease, Rare Variant hypothesis states that multiple rare variants with high penetrance contribute to the genetic susceptibility of common diseases (Pritchard, 2001; Schork et al., 2009). Figure 2 shows the features of common and rare variants distributed according to penetrance. Allele frequency and penetrance are inversely correlated for most of the disease-associated genetic variants (Figure 2). Very rare variants with small effects on the disease are difficult to identify, and common variants with large effects are rare in the population (Figure 2). Most of the common prostate cancer-associated variants [minor allele frequency (MAF) > 5%] with low to moderate effects have been identified through genome-wide association studies (GWASs), while prostate cancer family-based analyses, such as genetic linkage studies, have been more suitable for identifying low-frequency (MAF 1-5%) and rare susceptibility variants (MAF < 1%) with high penetrance (Demichelis and Stanford, 2015). Approximately 30% of the prostate cancer risk is estimated to be explained by common variants while rare and low-frequency variants contribute to approximately 5% of the disease risk, which indicates that the genetic susceptibility of prostate cancer remains largely unexplained (Attard et al., 2016; Demichelis and Stanford, 2015) (Figure 3).

The contribution of genetic factors to the risk of prostate cancer is evident, but genetic susceptibility of aggressive prostate cancer is unclear. Somatic mutations are thought to drive disease progression, but emerging evidence indicates that germline variants may also affect this process (Isaacs, 2012). Prostate cancer-specific survival has been reported to be correlated between fathers and their sons, suggesting that heritability influences the prognosis of this disease (Hemminki et al., 2008; Lindstrom et al., 2007). In addition, the risk of fatal prostate cancer is increased in men with a family history of this type of cancer (Hemminki et al., 2011), and brothers of men diagnosed with high-grade prostate cancer are at an increased risk of high-grade cancer (Jansson et al., 2012). The familial concordance of prostate cancer survival may be due to host-related factors, such as health awareness and behaviours, leading to early diagnosis and treatment; however, the concordance in tumour characteristics observed indicates that a genetic component is at least partially involved (Jansson et al., 2012).
Figure 2. Allele frequency is inversely correlated with penetrance for most of the identified susceptibility variants. Picture was modified from McCarthy et al. (2008).

Figure 3. The genetic risk of prostate cancer is explained by both rare and common variants, but this risk remains largely unexplained. Picture was modified from Attard et al. (2016).
2.2.2 From linkage analyses to post-GWAS approaches

**Linkage analyses.** Linkage analyses identify genetic markers that co-segregate with a trait, i.e., a marker at a specific chromosomal locus and that trait are co-inherited. These analyses are performed to examine families containing multiple affected individuals. Subsequent screening of a target region identified through linkage is usually necessary to identify a candidate gene or other predisposing genetic factor contributing to disease susceptibility.

Several linkage analyses have been conducted to identify the chromosomal loci associated with predisposition to prostate cancer, however, the findings of these analyses, which map in 14 different chromosomes, have not been consistently replicated [reviewed in (Eeles et al., 2014)]. These discrepant results may be explained by, for example, differences in the cancer families analysed and the quality of genotypes and statistical methods used among studies, as well as the genetic heterogeneity of the disease and differences in susceptibility loci among populations studied (Schaid, 2004).

A few candidate genes have been identified through linkage and the subsequent analyses, including ribonuclease L (2',5'-oligoisoadenylate synthetase-dependent) (RNASEL) at 1q25 (Carpten et al., 2002), elaC ribonuclease Z 2 (ELAC2) at 17p11 (Tavtigian et al., 2001), macrophage scavenger receptor 1 (MSR1) at 8p22 (Xu et al., 2002) and homeobox B13 (HOXB13) at 17q21 (Ewing et al., 2012). However, inconsistent and inconclusive findings have been reported for RNASEL, ELAC2 and MSR1 after their initial discoveries [reviewed in (Alvarez-Cubero et al., 2013)], whereas the results for HOXB13 have been replicated in multiple studies. Meta-analysis has reported that the low-frequency G84E mutation in HOXB13 (frequency of 0.1-4.9% in affected men) increases the risk of prostate cancer by four-fold and is particularly associated with early-onset and familial prostate cancer, in addition to aggressive disease (Huang and Cai, 2014). The frequency of this mutation is highest in Northern Europe (1.06%), and it has been suggested to have originated in Finland around the turn of the 19th century, prior to its spread to other geographic regions due to Finnish population migration (Chen et al., 2013). A Finnish population-based study has identified an increased risk of prostate cancer in G84E mutation carriers, particularly for familial and early-onset disease, which is consistent with the findings of other studies (Laitinen et al., 2013). A total of four genetic tests that include analysis of HOXB13 are offered by CLIA-certified laboratories to be used in risk assessment, screening, diagnosis, mutation confirmation and therapeutic management of prostate cancer; however, the clinical utilities of these tests, i.e., how likely they are
to significantly improve patient outcomes, have not been established (Genetic Testing Registry; Rubinstein et al., 2013).

Over ten genome-wide linkage analyses have been performed to search for genomic regions linked to aggressive prostate cancer [reviewed in (Isaacs, 2012)]. These studies have aimed to reduce the challenges associated with locus and disease heterogeneity by targeting the aggressive form of this disease (Isaacs, 2012). Aggressive cancer was defined in these studies using Gleason grade or using multiple indicators such as stage and grade of cancer, PSA level and prostate cancer as a cause of death (Isaacs, 2012). Linkage peaks have been observed in 12 different chromosomes, and the genomic regions, in which linkage has been replicated in multiple studies, are located in 5q, 7q, 19q and 22q [reviewed in (Isaacs, 2012)].

**Association studies.** Case-control-based studies search for an association between a marker and disease by comparing the frequency of the marker between cases (men diagnosed with disease) and controls (men without disease diagnosis). Association studies including a large number of individuals are suitable for identifying common risk variants with small effects on a particular disease (Stranger et al., 2011). A candidate gene study targets specific regions of a gene to identify disease-associated variants. However, the early association studies based on a limited number of variants in suspected cancer-associated genes were largely unsuccessful in detecting robust associations (Varghese and Easton, 2010). An approach for analysing hundreds of thousands variants simultaneously is the genotyping of single-nucleotide polymorphisms (SNPs) within the genome in a GWAS (Varghese and Easton, 2010). Detection of the nonrandom association of alleles at two or more loci, i.e., linkage disequilibrium (LD) of SNPs (Slatkin, 2008), enables the testing of only a proportion of the variants in the genome (Stranger et al., 2011). The representative SNPs in LD with other nearby SNPs are called tagging SNPs (tag SNPs) (Halperin and Stephan, 2009a). Based on a recent comparison of commercially available SNP chips, SNP panels for GWASs include 600,000-2.5 million SNPs targeting variants with MAFs > 1.5% that are estimated to cover < 50% of genome-wide variation (Ha et al., 2014). This coverage is much lower than that advertised by the manufacturers, which might be attributed to inconsistent determination of coverage or to the reference population used (Ha et al., 2014). Many GWASs have been performed since their advent in the year 2005 (Wang et al., 2015) (Figure 4), and these studies have identified a large number of cancer-associated loci that were not previously suspected to contribute to cancer development (Varghese and Easton, 2010). The challenges associated with performing GWASs include the following: false-positive results may be obtained due
to the large amount of SNPs tested; inclusion of a mixture of individuals of different ancestries may confound the results; rare disease variants will likely be missed; it is difficult to interpret the results because a considerable number of disease-associated SNPs are located in non-coding regions; and identification of causal variants requires further study (Stranger et al., 2011; Wang et al., 2015).

A total of 22 GWASs on prostate cancer were available in the GWAS Catalog provided by the National Human Genome Research Institute and the European Bioinformatics Institute by November 13, 2015 (GWAS Catalog; Welter et al., 2014). Approximately 100 common prostate cancer risk SNPs located on chromosomes 1-14, 16-22 and X have been identified through GWASs, and these SNPs account for 33% of the familial prostate cancer risk in European-ancestry populations (Al Olama et al., 2014). The prostate cancer risk for men in the top 10% of the polygenic risk distribution based on 100 markers is increased by 2.9-fold and that for men in the top 1% is increased by 5.7-fold compared with the population average suggesting that a screening method based on genetic risk profiling may be useful for reducing the overdiagnosis of prostate cancer; however, further studies on this matter are needed (Al Olama et al., 2014).

Few GWASs have searched for SNPs predisposing to aggressive prostate cancer or prostate cancer mortality. A SNP located in chromosome 19q13 has been demonstrated to predispose men to aggressive prostate cancer in meta-analysis of four GWASs (Amin Al Olama et al., 2013), and SNPs in 3q26, 5q14, 10q26, 15q21 and 19q13 have been observed to be associated with prostate cancer aggressiveness in GWASs not included in meta-analysis (Berndt et al., 2015; Nam et al., 2011). Furthermore, some evidence of association with aggressive disease has been reported for 15q13 (P < 1 x 10^{-4}) (FitzGerald et al., 2011), and 2q31.2, 11q12.2 and 11q14.1 have been identified to show association with prostate cancer mortality (P < 1 x 10^{-5}) (Penney et al., 2010), although no genome-wide statistical significances were observed (P < 1 x 10^{-7}).
Figure 4. The number of studies of human traits using linkage analyses, genome-wide association analyses and next-generation sequencing techniques published from 1980-2014 (Wang et al., 2015). This image was obtained and modified from “A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing”, Wang, Lu and Zhao, 2015, http://dx.doi.org/10.3389/fgene.2015.00149. It is licensed under Attribution CC BY, http://creativecommons.org/licenses/by/4.0/.

**Post-GWAS approaches.** GWASs identify associations between genomic regions and diseases rather than directly detecting causal variants. A region of interest can be scanned for disease-associated variants, for example, using imputation or targeted next-generation sequencing (NGS). Imputation methods statistically define ungenotyped SNPs based on genotyped tag SNPs using a known reference panel of variants and haplotypes (Halperin and Stephan, 2009b). The human genome consists of haplotype blocks, i.e., regions over which the historical recombination rate is very low (Gabriel et al., 2002). As a consequence, the alleles at different loci in a haplotype block are inherited together as a haplotype. Alleles of tag SNPs can be used to predict those at other loci in the same haplotype (Halperin and Stephan, 2009b). The 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) and the International HapMap Project (International HapMap 3 Consortium et al., 2010) have catalogued a comprehensive array of common human genetic variants that can be used as a reference panel to impute ungenotyped SNPs. The aim of the completed 1000 Genomes Project was to identify most of the human genetic variants with a frequency of > 1% and to provide accurate haplotype information for multiple human populations (1000 Genomes Project Consortium et al., 2010). A total of 88 million variants, including SNPs, short insertions and deletions and other structural variants, in populations of various ancestries were characterized (1000 Genomes Project Consortium et al., 2015). The objective of the International HapMap Project was to produce a haplotype map (HapMap) of the human genome that could be used to study the origins of diseases (International HapMap Consortium, 2003). HapMap
phases I, II and III have described approximately 1.3-4 million variants in a total of 11 populations (International HapMap Project).

The development of NGS technologies has provided cheaper and higher-throughput alternatives for identification of all SNPs within a target region or within the whole exome or genome compared to traditional Sanger sequencing (Grada and Weinbrecht, 2013). In addition to SNPs, copy number changes, translocations, and inversions, as well as epigenetic variations, i.e., modifications of DNA or chromatin in the absence of changes in the DNA sequence, can be detected using NGS (Wang et al., 2015). Furthermore, the entire transcriptome, i.e., RNA molecules, can be sequenced using NGS technologies as an alternative to microarrays in gene expression studies (Grada and Weinbrecht, 2013). NGS is based on massively parallel sequencing, which concurrently produces thousands or millions of short sequence reads that are aligned against reference sequences (Grada and Weinbrecht, 2013). Compared to whole-exome and genome sequencing, targeted sequencing focusing on a specific region of the genome is more affordable and less time-consuming providing deeper coverage for detection of low-frequency and rare variants (Grada and Weinbrecht, 2013; Xuan et al., 2013). In addition, targeting an analysis to a specific region allows for examination of larger sample sets, which increases the power (Wang et al., 2015). The objective of targeted sequencing is to identify disease-causing variants that are in LD with the associated SNPs identified, for example, through a GWAS (Freedman et al., 2011). An increasing number of studies are exploiting NGS technologies to better understand the origins of diseases (Figure 4), for example, using study designs that were popular prior to GWASs, such as candidate gene and linkage analyses (Wang et al., 2015). The major challenges associated with NGS technologies are related to the amount of data generated; data storage, analysis and interpretation (Wang et al., 2015).

Most disease and trait-associated SNPs (~93%) identified through GWASs are located within non-coding sequences, indicating involvement in gene regulation of at least for a proportion of these variants (Maurano et al., 2012). Therefore, one of the major focuses of studies in the post-GWAS era has been characterization the impacts of disease-associated variants in cells. Rather than directly conducting time-consuming functional analyses, the available public data on functional elements reported by major collaborative projects can be used to predict functionality and thereby to prioritize variants for functional assays. The Encyclopedia of DNA Elements (ENCODE) Project, which aimed to identify all functional elements in the human genome (ENCODE Project Consortium, 2004), produced comprehensive genome-wide information related to genome activity in multiple human cell types.
(ENCODE Project Consortium et al., 2012), whereas the Roadmap Epigenome Project focused on the generation of genome-wide epigenetic maps for human primary cells and tissues to better understand the relationships between epigenetic mechanisms and human health and disease (Bernstein et al., 2010). The Genotype-Tissue Expression (GTEx) Project was launched to develop a data and sample resource for researchers to study the effects of genetic variants on gene expression in multiple human tissue types (GTEx Consortium, 2013). Web tools, such as HaploReg (Ward and Kellis, 2012) and RegulomeDB (Boyle et al., 2012), are available for functional annotations, especially of non-coding genomic variants, based on the data produced by the aforementioned projects. These annotations provide information on chromatin structure, binding sites of transcription factors, epigenetic modifications of histones, and active and inactive chromosome states, as well as information on the effects of SNPs on gene expression (Boyle et al., 2012; Ward and Kellis, 2012).

Complex trait-associated variants identified through GWASs, including prostate cancer risk SNPs (Jiang et al., 2014), are more likely to affect gene expression levels, i.e., to act as expression quantitative trait loci (eQTLs), compared with other randomly chosen MAF-matched SNPs from GWAS platforms, suggesting that risk SNPs frequently contribute to traits by altering the level or timing of protein expression rather than by merely affecting protein structure (Nicolae et al., 2010). Regulatory SNPs may affect the expression of a nearby gene as cis-acting variants, or they have a more distant trans-effect on a gene located on the same or another chromosome (Nica and Dermitzakis, 2013). Studies have mainly focused on determining the cis-effects of disease-associated SNPs, at least partly due to the heavy computational burden related to assessing the whole genome for potential regulatory effects. Common, low-frequency and rare variants contribute to cis-effects, but additional factors with influences other than cis-effects have been suggested to account for over half of the total heritability of gene expression (Grundberg et al., 2012). However, the identification of trans-acting variants requires large sample sizes, and the tissue-dependency of trans-effects adds complexity to studies of such variants (Grundberg et al., 2012). In general, varying proportions (approximately 10-70%) of overlap in SNP-regulated gene expression have been found between tissues (Dimas et al., 2009; Fu et al., 2012; GTEx Consortium, 2015; Nica et al., 2011), with more similar cell types sharing higher numbers of eQTLs (Brown et al., 2013). However, trait-associated SNPs in particular have been found to exert tissue-dependent effects (Fu et al., 2012) emphasizing the importance of examining data from multiple tissue types relevant to a phenotype.
Regulatory effects of prostate cancer risk SNPs in prostate tumours were reported by three recent studies examining expression data from the Cancer Genome Atlas (n = 145 prostate tumours) (Amin Al Olama et al., 2015; Han et al., 2015b; Li et al., 2014). Multiple overlapping eQTLs were identified among the three studies, with the most significant associations observed at 5p15 with *iroquois homeobox 4* (*IRX4*), at 6q25 with *regulator of G-protein signaling 17* (*RGS17*) and at 17p13 with *family with sequence similarity 57 member A* (*FAM57A*) and *vacuolar protein sorting 53 homolog (S. cerevisiae)* (*VPS53*). In addition, an independent study investigating the effects of GWAS-identified prostate cancer loci on gene expression in prostate tumours has reported the strongest association at 5p15 with *IRX4* (Xu et al., 2014). Overall, 18-45% of the analysed risk loci have been demonstrated to contain eQTLs acting in cis in the aforementioned studies. In addition to prostate tumours, the regulatory potentials of prostate cancer GWAS SNPs have been studied in lymphoblastoid cell lines (LCLs), prostate cancer stroma, normal prostate tissues and adipose and skin tissues. Among the most significant associated eQTLs detected in prostate tumours, *IRX4* was also identified in skin tissue, *RGS17* was detected in LCLs, and *VPS53* was found in skin and adipose tissues (Amin Al Olama et al., 2015). LCL-specific enrichment of cis-eQTL signals has been reported for prostate cancer-associated SNPs in the Caucasian population (Jiang et al., 2014). Specifically two SNPs in the chromosomal regions 19p13 and 12q24 were demonstrated to function as cis-eQTLs in LCLs and to be located in transcription factor-binding sites (Jiang et al., 2014). A total of eight risk SNPs in the chromosomal regions 2p21, 2q31, 5p15, 8q24, 11q13, 17q24 and 22q13 have been shown to exhibit trans-effects on one or multiple genes in prostate cancer stroma, with a total of 47 SNP-gene associations (Chen et al., 2015). The highest number of associated genes (n = 32) has been reported for rs10896449 at 11q13.3 (Chen et al., 2015). In normal prostate tissues, 51 out of 100 risk regions were identified to contain cis-eQTLs, with 88 associated genes (Thibodeau et al., 2015). These cis-eQTLs were located in chromosomes 1-7, 10-12, 14, 16, 17, 19-22 and X (Thibodeau et al., 2015).

2.2.3 11q13-14

One of the chromosomal regions exhibiting linkage in Finnish prostate cancer families is 11q14. (Schleutker et al., 2003). Linkage between 11q13.4 (D11S1314) and
11q22.1 (D11S898) has been detected in five families originating from the Western coastal area of Finland, but these families did not share any particular clinical characteristics that differentiated them from the other families in the study (Schleutker et al., 2003). A linkage study conducted by the International Consortium for Prostate Cancer Genetics (ICPCG) on 1233 prostate cancer families has reported replication of the signal at 11q13.4 (D11S1314) (Christensen et al., 2010). However, the findings of these two studies were not independent, as the Finnish families examined in the study performed by Schleutker et al. were included in analysis conducted by ICPCG. Fine mapping of the 11q14 linkage peak in Finnish prostate cancer families did not result in identification of a stronger marker than those reported in the original study (Rokman et al., 2005). In addition, chromosomal region 11q14 was identified in a linkage study of families with prostate and colon cancer; however, no genome-wide statistical significance was observed (Fitzgerald et al., 2010). Further, no linkage in this region was detected in Finnish families with both prostate and gastric cancer (n = 3) (Schleutker et al., 2003).

In addition to prostate cancer families in general, the locus 11q14.1-14.3 has shown linkage in families with aggressive prostate cancer according to a large pooled ICPCG linkage study that searched for aggressive prostate cancer-linked chromosomal loci in 166 families, including the four Finnish families from the genome-wide linkage study that originally identified 11q14 (Schaid et al., 2006). The families included in the ICPCG study had three or more men with aggressive prostate cancer based on the following criteria developed by the ICPCG Epidemiology Subcommittee: a regional or distant disease stage; a Gleason grade at diagnosis of ≥7; a poorly differentiated grade if the Gleason grade was not available; pretreatment PSA at diagnosis of ≥ 20 ng/ml; or if deceased, death from metastatic prostate cancer before 65 years of age (Schaid et al., 2006). The strongest linkage of the 11q region was observed among families with an average age at diagnosis of 65 years or less (Schaid et al., 2006).

Table 2 summarizes the prostate cancer-associated SNPs in 11q13.3 identified in GWASs and in subsequent replication and fine mapping studies. An association of 11q13.3 was observed simultaneously by two GWASs reporting the association of two high-LD SNPs with prostate cancer: rs7931342 (Eeles et al., 2008) and rs10896449 (Thomas et al., 2008). This association has been replicated in multiple studies (Agalliu et al., 2013; Amin Al Olama et al., 2015; Chang et al., 2011; Chung et al., 2011; Gudmundsson et al., 2009; Hooker et al., 2010; Kote-Jarai et al., 2008; Schumacher et al., 2011; Waters et al., 2009; Zheng et al., 2009), of which few included familial patients (Breyer et al., 2009; Eeles et al., 2008; Jin et al., 2012;
Teerlink et al., 2014). In addition to the aforementioned signal, two other independent association signals at 11q13.3 have been detected and replicated in subsequent studies (Agalliu et al., 2013; Amin Al Olama et al., 2015; Chung et al., 2011; Gudmundsson et al., 2009; Zheng et al., 2009). Multiple candidate variants (n = 5-30) for these individual signals have been suggested for further analysis (Amin Al Olama et al., 2015; Chung et al., 2012). Most GWASs and replication and fine mapping studies have not found an association of the 11q13.3 region with aggressive prostate cancer (Agalliu et al., 2013; Gudmundsson et al., 2009; Hooker et al., 2010; Kote-Jarai et al., 2008; Teerlink et al., 2014; Thomas et al., 2008; Waters et al., 2009; Zheng et al., 2009). However, a GWAS focusing on advanced disease identified rs7940107, which is a surrogate of rs10896449 (signal III) (Schumacher et al., 2011); in addition, rs11228586 (signal II) has been demonstrated to be associated with unfavourable pathological characteristics of prostate cancer (McGuire et al., 2012). Further, the SNPs rs7931342 (Lange et al., 2012) and rs10896449 (signal III) (Agalliu et al., 2013; Jin et al., 2012) have been associated with a young age at diagnosis.

Most of the identified SNPs in 11q13.3 are located within an intergenic region, except for rs12418451, which is situated within long intergenic non-coding RNA (lincRNA) (Ensembl genome browser; Flicek et al., 2014), suggesting that the SNPs may affect gene regulation as a prostate cancer-predisposing mechanism. Such effects of the SNPs have been identified in LCLs (Amin Al Olama et al., 2015) and prostate tumour stroma (Chen et al., 2015) but not in prostate tumours or skin, adipose (Amin Al Olama et al., 2015) or normal prostate tissues (Thibodeau et al., 2015). The SNP rs10792032 (signal III) and variants in LD with the lead SNP have been shown to have cis-effects on the expression of cyclin D1 (CCND1), myeloma overexpressed (MYEOV) and oral cancer overexpressed 1 (ORA0V1) in LCLs (Amin Al Olama et al., 2015), and rs10896449 (signal III) has been identified to have a trans-effect on a total of 32 genes located in 17 different chromosomes in prostate tumour stroma (Chen et al., 2015).

The chromosomal region 11q13 has also been detected in other cancers. Among the prostate cancer risk variants, the high-LD SNPs rs7931342 and rs10896449 have been demonstrated to be associated with risk and family history of breast cancer, respectively (Agalliu et al., 2013; Song et al., 2008), but they have not been detected in breast cancer GWASs [GWAS hits reviewed in (Fachal and Dunning, 2015)]; instead, two other SNPs at 11q13.1 and 11q13.3 have been shown to be associated with breast cancer (Michaillidou et al., 2013; Turnbull et al., 2010). Locus 11q13.3 harbours a total of three independent association signals, which together explain approximately 2% of the familial breast cancer risk in Europeans (French et al.,
The associated region partly overlaps with the 11q13 region identified in prostate cancer. The breast cancer-associated variants affect enhancer and silencer elements and consequently alter the binding of specific transcription factors that likely regulate CCND1 expression (French et al., 2013). In addition to breast cancer, other SNPs at 11q13.3 and 11q13.4 are associated with renal cell carcinoma and colorectal cancer, respectively (Dunlop et al., 2012; Purdue et al., 2011). Interestingly, CCND1 has been suggested to be a target gene in renal cell carcinoma; a long-range enhancer of this gene overlaps with the 11q13.3 susceptibility locus and binds a specific transcription factor of the gene, and this process is modulated by the SNPs on the enhancer (Schodel et al., 2012).
Table 2. Prostate cancer-associated SNPs at 11q13.3 (Chr11:68906570-69453985, build 37).

<table>
<thead>
<tr>
<th>Signal/SNP(^a)</th>
<th>Study</th>
<th>Study type</th>
<th>Ethnic origin</th>
<th>Study size (ca/co)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent signal I</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10896438</td>
<td>Chung et al. 2011</td>
<td>Fine mapping</td>
<td>European</td>
<td>10272/9123</td>
<td>9.98 x 10(^{-11})</td>
</tr>
<tr>
<td>rs12418451</td>
<td>Zheng et al. 2009</td>
<td>Fine mapping</td>
<td>Caucasian</td>
<td>7012/4775</td>
<td>1.2 x 10(^{-4})</td>
</tr>
<tr>
<td><strong>Independent signal II</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12793759</td>
<td>Chung et al. 2011</td>
<td>Fine mapping</td>
<td>European</td>
<td>10272/9123</td>
<td>1.65 x 10(^{-13})</td>
</tr>
<tr>
<td>rs11228586</td>
<td>Gudmundsson et al. 2009</td>
<td>Fine mapping</td>
<td>European</td>
<td>7301/7053</td>
<td>6.7 x 10(^{-12})</td>
</tr>
<tr>
<td><strong>Independent signal III</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10792032</td>
<td>Amin Al Olama et al. 2015</td>
<td>Fine mapping</td>
<td>European</td>
<td>25723/26274</td>
<td>3.5 x 10(^{-17})</td>
</tr>
<tr>
<td>rs7931342</td>
<td>Eeles et al. 2008</td>
<td>GWAS</td>
<td>European</td>
<td>5122/5260</td>
<td>1.7 x 10(^{-12})</td>
</tr>
<tr>
<td>Kote-Jarai et al. 2008</td>
<td>Replication</td>
<td>Multiethnic</td>
<td>7370/5742</td>
<td>3 x 10(^{-11})</td>
<td></td>
</tr>
<tr>
<td>Waters et al. 2009</td>
<td>Replication</td>
<td>Multiethnic</td>
<td>2768/2359</td>
<td>8.4 x 10(^{-3})</td>
<td></td>
</tr>
<tr>
<td>Chang et al. 2011</td>
<td>Replication</td>
<td>African</td>
<td>2445/2018</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>Thomas et al. 2008</td>
<td>GWAS</td>
<td>European</td>
<td>4760/5133</td>
<td>1.76 x 10(^{-9})</td>
<td></td>
</tr>
<tr>
<td>Jin et al. 2012</td>
<td>Replication</td>
<td>European</td>
<td>57301979c</td>
<td>2.31 x 10(^{-3})</td>
<td></td>
</tr>
<tr>
<td>Hooker et al. 2010</td>
<td>Replication</td>
<td>African American</td>
<td>454/301</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>Chang et al. 2011</td>
<td>Replication</td>
<td>African</td>
<td>2056/1898</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td>Agalliu et al. 2013</td>
<td>Replication</td>
<td>Ashkenazic</td>
<td>979/1251</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Teerlink et al. 2014</td>
<td>Replication</td>
<td>Multiethnic</td>
<td>9560/2283c</td>
<td>4.8 x 10(^{4})</td>
<td></td>
</tr>
<tr>
<td>Zheng et al. 2009</td>
<td>Fine mapping</td>
<td>Caucasian</td>
<td>7012/4775</td>
<td>1.6 x 10(^{-11})</td>
<td></td>
</tr>
<tr>
<td>Chung et al. 2011</td>
<td>Fine mapping</td>
<td>European</td>
<td>10272/9123</td>
<td>7.94 x 10(^{-19})</td>
<td></td>
</tr>
<tr>
<td>rs10896450</td>
<td>Breyer et al. 2009</td>
<td>Replication</td>
<td>Caucasian</td>
<td>523/623</td>
<td>2.2 x 10(^{-4})</td>
</tr>
<tr>
<td>Gudmundsson et al. 2009</td>
<td>Fine mapping</td>
<td>European</td>
<td>7301/7053</td>
<td>2.6 x 10(^{-8})</td>
<td></td>
</tr>
<tr>
<td>rs11228594</td>
<td>Schumacher et al. 2011</td>
<td>GWAS</td>
<td>European</td>
<td>10999/11190</td>
<td>1.2 x 10(^{-9})</td>
</tr>
<tr>
<td>rs7940107</td>
<td>Schumacher et al. 2011</td>
<td>GWAS</td>
<td>European</td>
<td>5444/11190</td>
<td>4.7 x 10(^{-8})</td>
</tr>
</tbody>
</table>

\(^a\) SNPs representing an independent association signal are in strong LD with each other or were found to be correlated after conditioning on the lead SNP.

\(^b\) Signals I and II were replicated and refined by SNPs rs12275055 (P = 6.1 x 10\(^{-23}\)) and rs36225067 (P = 1.34 x 10\(^{-6}\)) (Amin Al Olama et al., 2015).

\(^c\) Affected men from prostate cancer families/total number of prostate cancer families

Abbreviations: ca = cases, co = controls, SNP = single-nucleotide polymorphism
2.3 EMSY

EMSY, BRCA2-interacting transcriptional repressor (EMSY) is located in a gene-dense region of chromosome 11q13.5 (Hughes-Davies et al., 2003). This gene is specific to the human genome, and no homologous genes have been identified in any other genomic region (Hughes-Davies et al., 2003); however, orthologues of this gene exist in genomes of 65 other species, including invertebrates (Ensembl genome browser; Flicek et al., 2014). Multiple splicing variants have been found to code for proteins ranging from 82 to 1337 amino acids in size (Ensembl genome browser; Flicek et al., 2014). The EMSY gene, encoding a 1322-amino acid protein, was identified in 2003 in a breast and ovarian cancer study (Hughes-Davies et al., 2003).

The structure of this protein suggests that it functions in chromatin-related processes. It contains an evolutionarily conserved EMSY N-terminal (ENT) domain (Hughes-Davies et al., 2003) composed of the first 100 encoded amino acids (Chavali et al., 2005). The α-helical bundle complex of the ENT domain exists as a homodimer in a crystal structure, indicating that EMSY is present in a dimeric form in cells, which could be important, for example, for its ability to bind other proteins (Chavali et al., 2005). The most conserved region of the ENT domain is similar in structure to the DNA-binding homeodomain of transcription factors, and the DNA-binding potential of this protein is further supported by the presence of a highly positively charged region on the dimer surface (Chavali et al., 2005).

Increasing evidence indicates that EMSY functions as a transcription factor. It has been shown to regulate the transcription of interferon-stimulated genes (ISGs) together with breast cancer 2 (BRCA2) (Ezell et al., 2012), transcription of growth and migration-related genes through a beta-catenin/T-cell factor (TCF) complex (Zhao et al., 2015) and transcription of an antimetastatic microRNA (Vire et al., 2014). In addition, EMSY interacts with the chromatin remodelling proteins chromobox 1 (CBX1) and zinc finger MYND-type containing 11 (ZMYND11) (Hughes-Davies et al., 2003) and with several protein complexes involved in transcriptional regulation (Garapaty et al., 2009; Malovannaya et al., 2011; Moshkin et al., 2009). The interaction of EMSY with BRCA2 was first identified in the breast and ovarian cancer study (Hughes-Davies et al., 2003) that originally described EMSY. The binding of the amino-terminal (N-terminal) region of EMSY to the transactivation domain of BRCA2 inhibits the transcriptional activity of BRCA2, which functions in both homologous recombinational DNA repair (Davies et al., 2001) and transcriptional activation (Milner et al., 1997). The function of EMSY in DNA repair is further supported by its colocalization with a DNA damage marker.
following DNA damage induction (Hughes-Davies et al., 2003) and by the observation that overexpression of its N-terminal region, including the BRCA2-interacting domain, induces BRCA2-deficient phenotypes in cells, such as structural chromosomal abnormalities (Raouf et al., 2005) and the decreased efficiency of homologous recombinational repair (Cousineau and Belmaaza, 2011). However, full-length EMSY may not have the same effect as the truncated protein on DNA repair (Raouf et al., 2005), as is supported by the lack of impairment in recombinational DNA repair observed in cancer cell lines in response to EMSY amplification (Wilkerson et al., 2011).

The contribution of EMSY to cancer has been evaluated mainly by copy number variation analyses, which have identified its amplification in breast (Bane et al., 2011; Brown et al., 2010; Hughes-Davies et al., 2003; Kirkegaard et al., 2008; Rodriguez et al., 2004), high-grade ovarian (Altinisik et al., 2011; Brown et al., 2006; Hughes-Davies et al., 2003) and pancreatic cancers (van Hattem et al., 2008). In addition to EMSY amplification, its deletion has been observed in breast tumours (Kirkegaard et al., 2008). Its amplification has been especially related to poor outcome of breast cancer patients (Brown et al., 2010; Hughes-Davies et al., 2003; Kirkegaard et al., 2008; Rodriguez et al., 2004), and EMSY copy number changes have also been identified in breast tumours in males (Kornegoor et al., 2012).

EMSY has been suggested to function as an oncogene because its overexpression promotes the growth and migration of ovarian (Zhao et al., 2015) and breast cancer cells, in addition to tumour formation and metastasis in vivo (Vire et al., 2014). At least two pathways have been identified by which EMSY affects tumour development. In ovarian cancer cells, EMSY increases expression of the oncogenes \textit{v}-myc avian myelocytomatosis viral oncogene homolog (MYC) and \textit{CCND1}, as well as that of \textit{axin} 2 (\textit{AXIN2}), through beta-catenin/TCF signalling (Zhao et al., 2015) and in breast cancer cells, it represses the expression of antimitastatic \textit{microRNA 31} (\textit{MIR31}) together with the transcription factor \textit{v}-ets avian erythroblastosis virus E26 oncogene homolog 1 (\textit{ETS1}) and histone demethylase lysine (K)-specific demethylase 5B (KDM5B) (Vire et al., 2014). In addition, EMSY has been connected to the AKT and Notch signalling pathways. It has been shown to function in a complex with BRCA2 and to repress ISGs in a v-akt murine thymoma viral oncogene homolog 1 (\textit{AKT1})-dependent pathway (Ezell et al., 2012). Interferons with antiviral and antitumour effects are secreted during the immune response, and they are exploited in cancer treatments (Dunn et al., 2006). However, the exact contribution of the AKT/BRCA2/EMSY/ISG pathway to the antitumour effects of interferons is not known (Ezell and Tsichlis, 2012), whereas the inhibition of viral
replication has been attributed to the activity of this pathway, which relieves EMSY repression (Ezell et al., 2012; Tian et al., 2015). Further, a complex of histone chaperones and other associated factors, including EMSY, has been shown to repress the transcription of Notch signalling pathway target genes (Moshkin et al., 2009). Interestingly, deregulation of Notch signalling has been detected during prostate cancer development (Carvalho et al., 2014), but the effect of EMSY aberrations on this pathway are unknown.

In summary, the findings of many studies have supported the contribution of EMSY to cancer development. Overexpression of EMSY has oncogenic effects, including promotion of tumour formation in vitro and in vivo. This overexpression is attributed to gene amplification in breast and ovarian cancers, but other mechanisms, such as altered gene regulation or dysfunction of the protein due to structural aberrations, might have similar consequences. In addition, EMSY functions in complex signalling pathways that are altered in cancer cells, and it may have a role in DNA repair. With regard to genetic susceptibility to cancer, the most interesting interacting protein of EMSY is BRCA2, which is encoded by a major breast cancer predisposition gene (Wooster et al., 1995). BRCA2 also increases susceptibility to prostate tumour development (Breast Cancer Linkage Consortium, 1999; Struweing et al., 1997), especially in men under 65 years of age (Breast Cancer Linkage Consortium, 1999; Kote-Jarai et al., 2011) with an absolute risk of 15% by age 65 (Kote-Jarai et al., 2011), and it has been connected to aggressive disease outcome and poor prognosis of prostate cancer (Edwards et al., 2010; Gallagher et al., 2010). The prostate cancer risk is increased in Finnish cancer families carrying BRCA2 mutations (Eerola et al., 2001), but no specific known BRCA2 mutations have been identified in Finnish prostate cancer patients (Ikonen et al., 2003), although the associations of other possible genetic variants of BRCA2 with prostate cancer risk in Finnish men has not yet been evaluated. The location of EMSY in the prostate cancer-linked chromosomal region, in which no candidate gene has been identified, and its contribution to other cancers make it a potential candidate gene that warrants further study in relation to prostate cancer susceptibility.
3 Aims of the Study

The general aim of this study was to obtain new knowledge regarding genetic susceptibility to prostate cancer, in particular, to the advanced form of this disease. The specific aims were as follows:

1. To identify genetic variants of *EMSY* at 11q13.5 and to study the associations of these variants with the risk of prostate cancer in Finnish men.

2. To further investigate the contributions of genetic variants and haplotypes at 11q13.5 to prostate cancer predisposition in the Finnish population.

3. To study the gene regulatory functions of the prostate cancer-predisposing genetic variants and haplotypes at 11q13.5.
4 Materials and Methods

4.1 Study subjects (I, II and III)

4.1.1 Samples

All samples used in the original publications (I, II and III) were obtained from prostate cancer patients (affected), males without prostate cancer (unaffected), and females of Finnish origin. Affected males from prostate cancer families, referred to here as familial cases, and unaffected male and female family members were from prostate cancer families from which samples were collected by the University of Tampere and Tampere University Hospital starting in 1995. A detailed description of the sample collection process is described elsewhere (Schleutker et al., 2000). Briefly, the following strategies were used for patient recruitment: questionnaires, advertisements in the media, information from a nation-wide cancer registry and church parish registries, and queries from urologists in Finland. All diagnoses were confirmed using the Finnish Cancer Registry or hospital-based patient records. The samples from the prostate cancer patients unselected for family history were collected by the University of Tampere and Tampere University Hospital in the Pirkanmaa Hospital District starting in 1996. Samples were also obtained from prostate cancer patients included in the screening arm of the ERSPC trial, which was initiated during the early 1990s. Detailed descriptions of the sample collection process and the screening protocol used for the Finnish component of the ERSPC PSA screening study are described elsewhere (Kilpelainen et al., 2010). Further, samples from male population controls were obtained from and collected by the Finnish Red Cross Blood Service. The anonymous controls were volunteer blood donors who were healthy at the time of blood donation. Individuals aged 18-65 years old were eligible to donate blood at the time that samples were obtained (Finnish Red Cross Blood Service). Freshly frozen specimens from primary and castration-resistant prostate cancer (CRPC) tumours were obtained by prostatectomy and transurethral resection of the prostate (TURP), respectively, from the Tampere University Hospital. The total amounts of the samples used in the original
publications are summarized in Table 3. The individuals from which the samples were collected overlapped among the studies as follows: 3113, 74, and 108 individuals overlapped between studies I and II, studies II and III, and studies I and III, respectively.

Full written informed consent concerning the samples and patient information was obtained from all participants. Permission to collect and use the samples and clinical information of the individuals was granted by the Ministry of Social Affairs and Health (licence 59/08/95), the Ethical Committee of the Tampere University Hospital (licences 95062 and R03203), the Institutional Review Board of the City of Tampere (licence 8595/403/2005) and VALVIRA (licence 9013/05.01.00.06/2010).

Table 3. Summary of samples used in studies I-III.

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Study I</th>
<th>Study II</th>
<th>Study III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected familial cases</td>
<td>188</td>
<td>-</td>
<td>105</td>
</tr>
<tr>
<td>Unaffected male family members</td>
<td>12</td>
<td>-</td>
<td>30</td>
</tr>
<tr>
<td>Female family members</td>
<td>3*</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>Affected unselected cases</td>
<td>2301</td>
<td>2716</td>
<td>-</td>
</tr>
<tr>
<td>Unaffected male family members</td>
<td>-</td>
<td>1318</td>
<td>-</td>
</tr>
<tr>
<td>Male population controls</td>
<td>923</td>
<td>908</td>
<td>-</td>
</tr>
<tr>
<td>LCL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected familial cases</td>
<td>-</td>
<td>-</td>
<td>24</td>
</tr>
<tr>
<td>Unaffected male family members</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>Affected unselected cases</td>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>Prostate tumour specimen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary tumours</td>
<td>-</td>
<td>-</td>
<td>71</td>
</tr>
<tr>
<td>CRPCs</td>
<td>-</td>
<td>-</td>
<td>11</td>
</tr>
</tbody>
</table>

* Includes one breast cancer case and one cancer case with undefined cancer.

b The Finnish component of the ERSPC trial.

Abbreviations: CRPC = castration-resistant prostate cancer, LCL = lymphoblastoid cell line, PSA = prostate-specific antigen

4.1.2 Classifying study subjects for association testing

The samples were divided into groups to examine the associations of genomic variants and haplotypes with prostate cancer and its clinicopathological features. Table 4 presents the classification criteria used.
Table 4. The classification criteria used to analyse associations with prostate cancer and its clinicopathological features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate cancer</td>
<td>prostate cancer diagnosis</td>
<td>no prostate cancer diagnosis (unaffected)</td>
<td>I, II, III</td>
</tr>
<tr>
<td>Cancer aggressiveness&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Gleason grade ≥ 7 or PSA value ≥ 20 ng/ml&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>Gleason grade &lt; 7 and/or PSA value &lt; 20 ng/ml&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Gleason grade ≥ 7 or PSA value ≥ 20 ng/ml&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>no prostate cancer diagnosis (unaffected)</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Gleason grade &lt; 7 and/or PSA value &lt; 20 ng/ml&lt;sup&gt;b,c&lt;/sup&gt;</td>
<td>no prostate cancer diagnosis (unaffected)</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Gleason grade ≥ 8&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Gleason grade ≤ 6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>II</td>
</tr>
<tr>
<td>Disease onset&lt;sup&gt;d&lt;/sup&gt;</td>
<td>diagnosis age ≤ 55 years</td>
<td>diagnosis age ≥ 56 years</td>
<td>II</td>
</tr>
<tr>
<td>PSA level&lt;sup&gt;a&lt;/sup&gt;</td>
<td>PSA level ≥ 20 ng/ml&lt;sup&gt;b&lt;/sup&gt;</td>
<td>PSA level ≤ 19 ng/ml&lt;sup&gt;b&lt;/sup&gt;</td>
<td>II</td>
</tr>
<tr>
<td>Cancer progression</td>
<td>biochemical recurrence present&lt;sup&gt;d&lt;/sup&gt;</td>
<td>no biochemical recurrence&lt;sup&gt;d&lt;/sup&gt;</td>
<td>II</td>
</tr>
<tr>
<td>Prostate cancer death</td>
<td>died of prostate cancer</td>
<td>prostate cancer diagnosis and follow up &gt; 6 years</td>
<td>II&lt;sup&gt;e&lt;/sup&gt;, III</td>
</tr>
</tbody>
</table>

<sup>a</sup> The feature was analysed as a quantitative trait in study III.
<sup>b</sup> At diagnosis
<sup>c</sup> Adapted from the ICPCG criteria (Schaid et al., 2006)
<sup>d</sup> Biochemical recurrence was defined as an elevated PSA level after prostatectomy, radiation therapy and hormone therapy.
<sup>e</sup> Deceased individuals were not included in Group 2.

Abbreviations: PSA = prostate-specific antigen
4.2 Laboratory methods (I, II and III)

The DNA and RNA extraction, variant screening and genotyping methods, as well as the methods used to determine the gene expression levels in the samples, are summarized in Table 5. All methods using commercial kits were conducted according to the manufacturers’ instructions. Most of the DNA and RNA from blood and tumour samples were extracted and the qualities of the extractions were assessed prior this study. LCLs were derived from peripheral mononuclear blood leucocytes by Epstein-Barr virus transformation and grown in BioWhittaker® RPMI 1640 medium (Lonza, Inc., Walkersville, MD, USA) with 1% l-glutamine (Sigma-Aldrich, St. Louis, MO, USA), 1% penicillin-streptomycin (Sigma-Aldrich) and 10% foetal bovine serum (Sigma-Aldrich).
Table 5. Laboratory methods used in studies I-III.

<table>
<thead>
<tr>
<th>Method (Supplier)</th>
<th>Blood</th>
<th>LCLs</th>
<th>Tumours</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DNA extraction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puregene Kit (Gentra Systems)</td>
<td>I, II, III</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Prep DNA/RNA MiniKit (Qiagen)</td>
<td></td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Phenol-chloroform technique</td>
<td></td>
<td></td>
<td>III</td>
</tr>
<tr>
<td><strong>RNA extraction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit (Life Technologies)</td>
<td>III</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAXgene Blood miRNA Kit (PreAnalytiX GmbH)</td>
<td>III</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNaseasy Mini Kit with DNase treatment (Qiagen)</td>
<td></td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>All Prep DNA/RNA MiniKit (Qiagen)</td>
<td>III</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRIzol® Reagent (Invitrogen)</td>
<td></td>
<td>III</td>
<td></td>
</tr>
<tr>
<td><strong>DNA and RNA quality assessment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ND-1000 spectrophotometer (NanoDrop Technologies)</td>
<td>I, II, III</td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td>Agilent 2100 Bioanalyzer (Agilent Technologies)</td>
<td>III</td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td><strong>Mutation screening</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger sequencing using ABI Prism® Sequencer 3130xl (Applied Biosystems)b</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mutation genotyping</strong> [number of genotyped variants]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TaqMan® SNP Genotyping Assay using ABI Prism® 7900HT (Applied Biosystems) [31]</td>
<td>I, II, III</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>KASPar® On Demand Assay (KBioscience) using ABI Prism® 7900HT [9]</td>
<td>II, III</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>HRMA using CFX384™ Real-Time System (Bio-Rad)c</td>
<td>I, II</td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td>Sanger sequencing using ABI Prism® Sequencer 3130xl[b]</td>
<td>I, II, III</td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td>Human OmniExpress BeadChip Kit (Illumina)d</td>
<td>III</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gene expression level determination</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library construction and RNA sequencinga</td>
<td></td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td>RT² PreAmp cDNA Synthesis Kit, Custom RT² Profiler PCR Array (Qiagen)f</td>
<td>III</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT² First Strand Kit, Custom RT² Profiler PCR Array (Qiagen)g</td>
<td>III</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Detailed description of suppliers: Agilent Technologies, Santa Clara, CA, USA; Applied Biosystems, Foster City, CA, USA; Bio-Rad Laboratories, Hercules, CA, USA; Gentra Systems, Inc., Minneapolis, MN, USA; Illumina, Inc., San Diego, CA, USA; Invitrogen, Carlsbad, CA, USA; KBioscience, Hoddesdon, Hertfordshire, UK; Life Technologies, Carlsbad, CA, USA; NanoDrop Technologies, Wilmington, DE, USA; PreAnalytiX GmbH, Feldbachstrasse, Switzerland; and Qiagen, Valencia, CA, USA

b Detailed description of Sanger sequencing is available in the original publication. PCR products were purified using rAPid Alkaline Phosphatase (Roche, Mannheim, Germany) and Exonuclease I (Fermentas, Vilnius, Lithuania) in studies II and III.

c Target sequences were amplified by PCR with fluorescent double-stranded DNA-binding dye (Sso Fast™ EvaGreen® Supermix, Bio-Rad) and then the amplified products were heated gradually from 65 to 95 ºC. As the DNA melted, fluorescent dye was released, resulting in generation of a melt curve for the amplicons. Mutation statuses were determined by Sanger sequencing.

d Performed prior this study by the Technology Centre, Institute of Molecular Medicine Finland, University of Helsinki, Finland.

e Whole-transcriptome sequencing of blood and tumour samples was performed prior to this study by the Beijing Genomics Institute (BGI HongKong Co., Ltd., Tai Po, Hong Kong) using Illumina HiSeq™ 2000 technology (Illumina). A detailed description is provided elsewhere (Annala et al., 2015; Siltanen et al., 2013).

f Performed by the Qiagen Service Core (Hilden, Germany).

g Abbreviations: HRMA = high-resolution melt analysis, LCL = lymphoblastoid cell line, PCR = polymerase chain reaction.
4.3 Computational methods

4.3.1 Tag SNP determination, imputation and association testing of imputed variants (II)

The screening of prostate cancer-associated variants located at 11q13.5 was conducted using imputation with tag SNPs. Haploblocks (n = 20) and candidate tag SNPs (n = 118) were determined using Haploview v4.2 (Barrett et al., 2005) with HapMap3 SNPs (Utah residents with ancestry from Northern and Western Europe) and 27 EMSY variants identified in study I. A total of 31 tag SNPs covering all haploblocks were selected for genotyping based on their associations with prostate cancer. Associations were analysed with imputed genotypes using BIMBAM (Servin and Stephens, 2007).

The imputation of 11q13.5 variants was conducted with the genotyped tag SNPs using IMPUTE2 v2.2.3 (Howie et al., 2011; Howie et al., 2009) and the 1000 Genomes reference panel (1000 Genomes phase 1 integrated variant set March 2012) (1000 Genomes Project Consortium et al., 2010). Imputed variants were included in association analysis based on the info scores of the rare (MAF < 1%, info score ≥ 0.9) and low-frequency and common variants (MAF ≥ 1%, info score ≥ 0.8). Variants deviating from Hardy-Weinberg equilibrium (HWE) in the controls were excluded from association analysis. The associations of 604 imputed variants according to five genetic models (additive, dominant, recessive, general and heterozygote) were tested using the Frequentist test with the score method implemented in SNPTEST v2.3.0 (Marchini and Howie, 2010).

4.3.2 Association testing of genotyped variants (I and II)

The genotyped variants were tested for HWE in the controls using PLINK v1.07 (PLINK; Purcell et al., 2007). The associations of the variants with prostate cancer and its clinicopathological features were assessed using Fisher’s exact test, the χ² test or the Cochran Armitage trend test with GraphPad Prism v5.02 for Windows (GraphPad Software, Inc., San Diego, CA, USA) or PLINK v1.07 (PLINK; Purcell et al., 2007). P values adjusted for correlated tests (P_{ACT}) (Conneely and Boehnke, 2007) were calculated to account for multiple testing using R v2.11.1 (R
Development Core Team, 2014). The results were considered significant when $P$ and $P_{ACT} < 0.05$.

### 4.3.3 Haplotype analysis and LD (I, II and III)

The LD between genomic variants, haplotype blocks and haplotypes was determined using Haplovview v4.1 and v4.2 (Barrett et al., 2005) for the genotyped EM3Y variants in study I ($n = 5$) and for the imputed and genotyped variants in study II ($n = 635$). Haploblocks were defined based on Gabriel’s definition (Gabriel et al., 2002). A significant association of a haplotype with prostate cancer was identified at a $P < 0.05$. Haplotypes of familial subjects from study III were determined using MERLIN v1.1.2 (Abecasis et al., 2002; Abecasis and Wigginton, 2005) and BEAGLE v3.3.2 (Browning and Browning, 2009) with a total of 222 markers (in-house genotyped variants and SNPs of the Human OmniExpress BeadChip). The identified haploblocks included flanking SNPs in LD ($r^2 \geq 0.5$) with the prostate cancer risk SNPs identified in studies I and II and other SNPs in LD with the flanking SNPs. LD was estimated using PLINK v1.07 (PLINK; Purcell et al., 2007). The associations of haplotypes with prostate cancer and its clinicopathological features were examined according to three genetic models (additive, dominant and recessive) using the family-based association test (Horvath et al., 2004). A 5% false discovery rate (FDR) (Benjamini and Hochberg, 1995) was used to detect significant associations. FDRs were calculated using R v3.1.0 (R Development Core Team, 2014).

### 4.3.4 eQTL analysis (III)

eQTL analysis was performed to examine the cis-effects of the nine prostate cancer risk variants (identified in I and II) on gene expression in prostate tumours, whole blood and LCLs. Expression levels were determined by RNA sequencing during the screening phase (prostate tumours and whole blood) and by real-time quantitative polymerase chain reaction (RT-qPCR) during the validation phase (prostate tumours and LCLs). The RNA sequencing reads were aligned with the Gencode19 gene set using TopHat2 v2.0.0 (Trapnell et al., 2009). The raw read counts were determined using HTSeq v0.5.3 (Anders et al., 2015) and were normalized using DESeq2 v1.22.0 (Anders and Huber, 2010). The high-expressing genes located within a 1-Mb window around the associated variants were tested for associations with the variants using an
in-house developed, two-sided generalized Mann-Whitney test (Fischer et al., 2014), with 1000 permutations and the R-package GeneticTools available from Comprehensive R Archive Network (R Development Core Team, 2014). Completely correlated SNPs ($r^2 = 1$) were grouped in tests of association. Correlations were calculated using PLINK v1.07 (PLINK; Purcell et al., 2007). Genes were selected for validation by RT-qPCR based on the uncorrected P values obtained during screening. The raw Ct-data generated from RT-qPCR was analysed according to the manufacturer’s protocol using RT² Profiler PCR Array Data Analysis v3.5 (PCR Array Data Analysis Web Portal). Expression levels were normalized to the best-suited housekeeping genes according to a multigroup analysis. High-expressing genes were tested for associations with genomic variants using the $2^{-\Delta\text{Ct}}$ values as described for the screening phase. The P values were corrected for multiple testing using a 5% FDR implemented in Benjamini-Hochberg method (Benjamini and Hochberg, 1995). The generalized Mann-Whitney test was shown to outperform a commonly used linear model in a power study of 80 samples.

In addition to genetic variants, the cis-effects of haplotypes on gene expression in whole blood were analysed. These associations were tested in a similar manner as the associations of haplotypes with prostate cancer and its clinicopathological features.

The Test Your Own eQTLs analysis tool of GTEx Portal v6 (GTEx Consortium, 2015; GTEx Web Portal) was used to test the associations of the prostate cancer risk SNPs with the gene expression in whole blood, LCLs and prostate tissues. A total of seven genes were chosen for testing from the single SNP and haplotype analyses using the genotype and expression data from the GTEx Portal. The results were considered significant at a P < 0.05.

### 4.3.5 Functional annotation

Functional annotation of the prostate cancer-associated variants was performed by in silico analyses. The following bioinformatics tools were used in the original publications: the integrated web server A Regulatory RNA Motifs and Elements Finder (RegRNA) v1 (Huang et al., 2006) (I); Human Splicing Finder v2.4.1 (Desmet et al., 2009) (II); and HaploReg v2 (Ward and Kellis, 2012) (II). Functional annotations of the variants rs72944758, rs200331695, rs10899221, rs12271056, rs143975731, rs12277366, rs2155225, rs2155222 and rs7126722 were re-analysed in this thesis using the updated version of HaploReg v4.1 (Ward and Kellis, 2012). This
analysis included annotations of open chromatin, bound proteins, chromatin states (15-state model) and altered regulatory motifs. The effects of rs72944758, rs200331695 and rs10899221 on splicing were re-predicted using the updated version of HumanSplicingFinder v3.0 (Desmet et al., 2009). The following splicing elements were included in analysis: potential splice sites (HumanSplicingFinder Matrices), potential branch points, enhancer motifs of serine/arginine (SR)-rich proteins and silencer motifs of a heterogeneous nuclear ribonucleoprotein. RegRNA analysis, which was used to predict the effect of rs200331695 in the original publication I, was not repeated because this SNP was included in the two other analyses.
5 Summary of the Results

5.1 Prostate cancer-predisposing genetic variation at 11q13.5

5.1.1 Genetic variants (I and II)

The screening of exons and exon-intron boundaries of the *EMSY* gene resulted in identification of 27 sequence variants, of which 16 were genotyped in the controls and tested for association with prostate cancer. The fine mapping of 11q13.5 resulted in detection of 604 imputed markers, which were tested for association with prostate cancer together with genotyped tag SNPs. A total of nine *EMSY* variants and six intergenic variants were chosen for validation from screening analyses; these variants were genotyped in a larger sample set and tested for associations (a maximum of 2301 unselected cases, the youngest affected patient from 184 prostate cancer families and a maximum of 923 controls in study I; and 2716 unselected cases, 1318 patients of the Finnish component of the ERSPC trial and 908 controls in study II). The occurrences of the validated prostate cancer-associated variants (n = 9) in the relevant sample sets are presented in Table 6.

A total of eight variants were found to increase the risk of prostate cancer (Table 7). The *EMSY* SNPs rs72944758 and rs10899221, which were in moderate LD with each other ($r^2 = 0.68$), were associated with unselected prostate cancer ($P_{ACT} = 0.008-0.03$), and a stronger association was observed in the PSA screening cases ($P_{ACT} = 0.002-0.003$) (Table 7). The high-LD ($r^2 = 0.85-0.99$) intergenic SNPs rs12277366, rs2155225, rs2155222 and the single-nucleotide insertion rs143975731 increased the risk of unselected prostate cancer ($P_{ACT} = 0.03$), but the association was replicated only for rs12277366 in the PSA screening cases ($P_{ACT} = 0.03$) (Table 7). Similarly, another pair of high-LD ($r^2 = 0.98$) intergenic SNPs, rs12271056 and rs7126722, exhibited an association only in the unselected cases ($P_{ACT} = 0.01-0.03$) and not in the PSA screening cases ($P_{ACT} > 0.05$) (Table 7).
Table 6. The occurrence of prostate cancer-associated variants in the sample sets (I and II).

<table>
<thead>
<tr>
<th>Variant</th>
<th>Familial cases</th>
<th>Unselected cases</th>
<th>PSA screening cases</th>
<th>Controls</th>
<th>Aggressive cases</th>
<th>Nonaggressive cases</th>
<th>Died of cancer</th>
<th>Cancer diagnosis and follow up &gt; 6 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs72944758 (A/G)</td>
<td>NA</td>
<td>2295/383/15</td>
<td>1139/162/6</td>
<td>734/151/9</td>
<td>316/60/1</td>
<td>845/136/9</td>
<td>355/62/3</td>
<td>976/148/7</td>
</tr>
<tr>
<td>rs200331695 (IVS6-43A&gt;G)</td>
<td>181/3/0</td>
<td>2097/16/0</td>
<td>NA</td>
<td>921/2/0</td>
<td>1000/13/0</td>
<td>856/1/0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rs10899221 (G/A)</td>
<td>NA</td>
<td>2151/507/36</td>
<td>1062/231/15</td>
<td>670/218/12</td>
<td>298/78/1</td>
<td>790/191/16</td>
<td>336/76/8</td>
<td>916/201/14</td>
</tr>
<tr>
<td>rs12271056 (T/C)</td>
<td>NA</td>
<td>1447/1036/182</td>
<td>676/537/87</td>
<td>432/393/70</td>
<td>201/136/33</td>
<td>524/400/57</td>
<td>206/161/44</td>
<td>628/427/66</td>
</tr>
<tr>
<td>rs12277366 (G/C)</td>
<td>NA</td>
<td>596/1328/768</td>
<td>289/657/359</td>
<td>174/434/293</td>
<td>70/185/120</td>
<td>230/497/264</td>
<td>68/197/151</td>
<td>260/573/300</td>
</tr>
<tr>
<td>rs2155222 (G/A)</td>
<td>NA</td>
<td>713/1314/648</td>
<td>326/663/311</td>
<td>205/446/245</td>
<td>87/188/98</td>
<td>272/493/223</td>
<td>92/199/119</td>
<td>307/566/254</td>
</tr>
<tr>
<td>rs7126722 (G/A)</td>
<td>NA</td>
<td>1457/1046/179</td>
<td>677/540/85</td>
<td>439/395/67</td>
<td>202/138/32</td>
<td>531/404/57</td>
<td>204/168/44</td>
<td>634/431/64</td>
</tr>
</tbody>
</table>

* The cases used to detect EMSY variants during the screening phase are excluded.

Abbreviations: NA = not analysed, PSA = prostate-specific antigen
The rare *EMSY* SNP rs200331695 (IVS6-43A>G) was not found to be associated with unselected prostate cancer ($P_{\text{ACT}} > 0.05$) (Table 7) but was found to increase the risk of familial prostate cancer [odds ratio (OR) = 7.5; CI 95% 1.3-45.5; $P_{\text{ACT}} = 0.02$]. This SNP was identified in the youngest affected patients of three prostate cancer families. Association of variants at 11q13.5 with familial prostate cancer was not assessed in study II. The prostate cancer-associated variants are presented in their genomic context in Figure 6a.

**Table 7.** Associations of genetic variants with prostate cancer.

<table>
<thead>
<tr>
<th>Variant (Tested allele)</th>
<th>Positiona</th>
<th>P(b) (%)</th>
<th>OR (95% CI)c</th>
<th>$P_{\text{ACT}}$</th>
<th>OR (95% CI)c</th>
<th>$P_{\text{ACT}}$</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs72944758 (A)</td>
<td>76173948</td>
<td>92.3</td>
<td>1.26 (1.04-1.52)</td>
<td><strong>0.03</strong></td>
<td>1.46 (1.17-1.83)</td>
<td><strong>0.003</strong></td>
<td>II</td>
</tr>
<tr>
<td>rs200331695 (IVS6-43A&gt;G) (G)</td>
<td>76174822</td>
<td>0.3</td>
<td>-</td>
<td>*</td>
<td>NA</td>
<td>NA</td>
<td>I</td>
</tr>
<tr>
<td>rs10899221 (G)</td>
<td>76181631</td>
<td>89.0</td>
<td>1.29 (1.10-1.52)</td>
<td><strong>0.008</strong></td>
<td>1.40 (1.16-1.69)</td>
<td><strong>0.002</strong></td>
<td>II</td>
</tr>
<tr>
<td>rs12271056 (T)</td>
<td>76349466</td>
<td>72.8</td>
<td>1.19 (1.06-1.34)</td>
<td>0.01</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td>rs143975731 (-)</td>
<td>76349716</td>
<td>46.2</td>
<td>1.14 (1.03-1.27)</td>
<td>0.03</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td>rs12277366 (G)</td>
<td>76350119</td>
<td>46.3</td>
<td>1.15 (1.03-1.28)</td>
<td><strong>0.03</strong></td>
<td>1.17 (1.04-1.32)</td>
<td><strong>0.03</strong></td>
<td>II</td>
</tr>
<tr>
<td>rs2155225 (A)</td>
<td>76351897</td>
<td>46.6</td>
<td>1.13 (1.01-1.26)</td>
<td>0.03</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td>rs2155222 (G)</td>
<td>76353805</td>
<td>50.4</td>
<td>1.15 (1.03-1.28)</td>
<td>0.03</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td>rs7126722 (G)</td>
<td>76354161</td>
<td>72.9</td>
<td>1.17 (1.04-1.32)</td>
<td>0.03</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
</tbody>
</table>

a Chromosome 11, build 37
b Population frequency
c OR and 95% CI are provided if $P_{\text{ACT}} < 0.05$.

* P$_{\text{ACT}}$ ≥ 0.05

Bold type signifies statistical significance.

Abbreviations: CI = confidence interval, F = frequency, NA = not analysed, OR = odds ratio, $P_{\text{ACT}}$ = P value adjusted for correlated tests.

The unselected and familial prostate cancer-associated variants were tested for association with the clinicopathological features of prostate cancer. The rare *EMSY* SNP rs200331695 increased the risk of aggressive prostate cancer ($P_{\text{ACT}} = 0.002$ and 0.03; case-case and case-control analyses, respectively) (Table 8). No association of rs200331695 with nonaggressive cancer was observed based on case-control analysis ($P > 0.05$). In addition, rs200331695 segregated with an aggressive disease outcome in two prostate cancer families, from which additional samples were available for study (Figure 5). The intergenic variants rs12271056, rs12277366, rs143975731, rs2155225, rs2155222 and rs7126722 were associated with prostate cancer death ($P_{\text{ACT}} = 2.4 \times 10^{-4}-0.02$) (Table 8). These associations were found to be inversely correlated with associations with unselected prostate cancer, i.e., the presence of the
unselected prostate cancer-predisposing alleles decreased the risk of cancer death and the presence of the nonpredisposing alleles increased this risk (Table 8).

**Table 8.** Associations of genomic variants with aggressive prostate cancer and prostate cancer death.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Variant</th>
<th>Allele</th>
<th>OR&lt;sup&gt;a,b&lt;/sup&gt;</th>
<th>95% CI&lt;sup&gt;a,b&lt;/sup&gt;</th>
<th>P&lt;sub&gt;ACT&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive cancer</td>
<td>rs72944758</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs200331695 (IVS6-43A&gt;G)</td>
<td>G</td>
<td>6.5/6.0&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.5-28.4/1.3-26.4&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.002/0.03&lt;sup&gt;c&lt;/sup&gt;</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>rs10899221</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs12271056</td>
<td>T</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs143975731</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs12277366</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs2155225</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs2155222</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs7126722</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td>Cancer death&lt;sup&gt;d&lt;/sup&gt;</td>
<td>rs72944758</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs200331695 (IVS6-43A&gt;G)</td>
<td>NA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>rs10899221</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs12271056</td>
<td>T (C)</td>
<td>0.76 (1.31)</td>
<td>0.64-0.91 (1.10-1.56)</td>
<td>0.01</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs143975731</td>
<td>G (C)</td>
<td>0.73 (1.37)</td>
<td>0.62-0.86 (1.17-1.61)</td>
<td>6.1 x 10^-4</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs12277366</td>
<td>G (C)</td>
<td>0.72 (1.40)</td>
<td>0.61-0.84 (1.19-1.64)</td>
<td>2.4 x 10^-4</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs2155225</td>
<td>A (T)</td>
<td>0.72 (1.38)</td>
<td>0.62-0.85 (1.18-1.63)</td>
<td>4.0 x 10^-4</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs2155222</td>
<td>G (A)</td>
<td>0.80 (1.25)</td>
<td>0.68-0.94 (1.07-1.47)</td>
<td>0.02</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>rs7126722</td>
<td>G (A)</td>
<td>0.74 (1.35)</td>
<td>0.62-0.88 (1.13-1.61)</td>
<td>0.003</td>
<td>II</td>
</tr>
</tbody>
</table>

<sup>a</sup> OR and 95% CI are provided if P<sub>ACT</sub> < 0.05.

<sup>b</sup> Case-case analysis unless otherwise specified: aggressive versus nonaggressive; died of cancer versus diagnosed with cancer with follow up > 6 years.

<sup>c</sup> Case-control (unaffected) analysis

<sup>d</sup> ORs and 95% CIs for both alleles are presented because of the inverse association of the prostate cancer non-risk allele with prostate cancer death.

<sup>*</sup>P<sub>ACT</sub> ≥ 0.05

Bold type signifies statistical significance.

Abbreviations: CI = confidence interval, NA = not analysed, OR = odds ratio, P<sub>ACT</sub> = P value adjusted for correlated tests.
Occurrence of the rare EMSY variant rs200331695 (IVS6-43A>G) with aggressive prostate cancer in families 359 and 017. The squares indicate males and the circles denote females. The black squares signify prostate cancer, and the circular symbols with black circles in the middle denote other types of cancer. A plus sign (+) indicates a mutation carrier (AG), and a minus sign (-) signifies wild type (AA). The current ages of the unaffected males in years (y) are presented below the symbols. Aggressive prostate cancer cases (A) were defined based on a Gleason grade ≥ 7 or PSA ≥ 20 ng/ml at diagnosis. Nonaggressive cases (NonA) were defined based on a Gleason grade < 7 and/or PSA < 20 ng/ml at diagnosis. The affection statuses, deceased subjects and current ages were updated on December 1st, 2015. In addition, the mutation statuses of the nonaggressive prostate cancer cases in family 017 were updated.
5.1.2 Haplotypes (I, II and III)

A total of 21 haplotype blocks were identified within 11q13.5, including a haploblock containing three genotyped EMSY variants (I), 16 haploblocks composed of both genotyped and imputed variants in 11q13.5 (II), and four haploblocks including the prostate cancer-associated variants identified in studies I and II or both prostate cancer-associated variants and other genotyped variants in 11q13.5 (III). Both prostate cancer risk increasing (n = 2) and decreasing haplotypes (n = 2) were observed (II). A haplotype including the prostate cancer risk alleles of the genotyped EMSY SNPs rs72944758 and rs10899221 and other imputed variants in and surrounding EMSY was found to increase the risk (P = 4.0 x 10^-4). The frequency of this haplotype in the population was 23.8%. In addition, the aforementioned alleles were included in a rare haplotype in the same haploblock (frequency of 1.6% in the population) that decreased the risk of prostate cancer (P = 0.01). Two haplotypes containing both genotyped and imputed intergenic variants were identified. A haplotype including the prostate cancer risk alleles of rs12271056, rs143975731 and rs12277366 increased the risk of prostate cancer (P = 0.02), while a haplotype containing the nonpredisposing alleles of rs2155225, rs2155222 and rs7126722 decreased this risk (P = 0.02). The population frequencies of the aforementioned haplotypes were 45.4% and 15.1%, respectively. The prostate cancer-associated haplotypes are presented in their genomic context in Figure 6a.

5.2 Functionality of prostate cancer risk variants and haplotypes

5.2.1 Functional annotation of variants

Functional elements coinciding with the prostate cancer risk SNPs rs72944758, rs200331695, rs10899221, rs12271056, rs143975731, rs12277366, rs2155225, rs2155222 and rs7126722 are presented in Figure 6b and 6c. Among the EMSY SNPs, rs10899221 exhibited the most evidence of functionality, as it was located within open chromatin and a binding site of a transcription factor CCAAT/enhancer binding protein beta (CEBPB), and it was predicted to alter a total of four regulatory motifs (Figure 6b). In addition, all EMSY variants were candidate SNPs affecting splicing because they were predicted to have effects on multiple splicing motifs (Figure 6b). All intergenic SNPs coincided with an enhancer element in T regulatory
cells, B cells, or natural killer cells of peripheral blood [abbreviated in Figure 6 as peripheral blood cells (PBCs)], in addition to other cell types (Figure 6c). Furthermore, all intergenic SNPs, except for rs2155225, were predicted to modify one or several regulatory motifs (Figure 6c).
Figure 6. The functionality of prostate cancer risk variants at 11q13.5. (A) The prostate cancer-associated variants, haplotypes and eQTLs in 11q13.5 in their genomic context (build 37). The prostate cancer risk SNPs, haplotypes and eQTLs are indicated in the figure. In addition to RefSeq genes, lncRNAs in chr11:76,330-76,380 kb are presented in blue colour. The eQTLs and eQTL-associated genes are marked in red colour. The rectangles denote exons. (B and C) Functional elements coinciding with prostate cancer risk variants. The number of cell types in which open chromatin or enhancer elements were detected is indicated in round brackets. If a functional element coincided with a variant in peripheral blood cells (PBCs) (primary T cells, B cells or natural killer cells) or in a LCL, it is specified in the figure. The numbers of changed regulatory motifs are provided in brackets. The crosses indicate altered splicing elements. Moderate or high LD between the variants is indicated by colours.
5.2.2 Regulation of gene expression (III)

5.2.2.1 Genetic variants

The cis-effects of the prostate cancer risk SNPs on gene expression were analysed in whole blood, prostate tumours and LCLs. In addition, the GTEx Portal data for whole blood, LCLs and prostate tissues were used for validation and to study the regulatory potentials of the SNPs in prostate tissues. The intergenic SNPs affected the expression of diacylglycerol O-acyltransferase 2 (DGAT2) in prostate tumours and that of AP001189.4 in whole blood (Table 9). The prostate cancer death-associated risk alleles of rs143975731 (G), rs12277366 (C), rs2155225 (T) and rs2155222 (A) decreased the expression of DGAT2 ($P = 0.035-0.043$ and FDR = 0.025-0.045; screening and validation set, respectively) (Table 9) (Figure 7). The effects of the SNPs on DGAT2 expression were stronger in the validation set of tumours (Table 9). The regulatory effects of the intergenic SNPs on AP001189.4 expression were not examined in LCLs, which were used to validate the effects observed in whole blood, due to weak expression of this gene, but strong associations were observed in whole blood using the GTEx Portal (Table 9). Among the SNPs listed in the GTEx Portal, rs12277366 exhibited the strongest association ($P = 8.7 \times 10^{-5}$), and the association of this SNP with AP001189.4 expression was also identified in our analysis ($P = 0.03$) (Table 9). The prostate cancer death-associated risk allele of rs12277366 (C) increased AP001189.4 expression (Figure 7). The identified and replicated eQTLs and their target genes are presented in Figure 7.

In addition to the intergenic SNPs, DGAT2 expression was observed to be associated with rs200331695 in LCLs (FDR = 0.048). However, replication analysis using the GTEx Portal data for LCLs could not be conducted because this SNP was missing from the Portal. Similarly, effects of SNPs on CAPN5 (rs12277366, rs2155222, rs2155225), EMSY (rs72944758, rs10899221) and PRKRIR (rs72944758, rs10899221, rs12271056, rs12277366, rs2155222) in prostate tissues were observed and analysed using the GTEx Portal only ($P < 0.05$).
Table 9. The effects of prostate cancer risk SNPs on gene expression in prostate tumours and whole blood.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Gene, Variant</th>
<th>Screening set, P</th>
<th>Replication set, P (FDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate tumour</td>
<td>DGAT2</td>
<td>primary tumours and CRPCs</td>
<td>primary tumours</td>
</tr>
<tr>
<td>rs72944758</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>rs200331695 (IVS6-43A&gt;G)</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>rs10899221</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>rs12271056, rs7126722&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.045</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>rs143975731, rs12277366, rs2155225&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.043</td>
<td>0.018 (0.045)</td>
<td></td>
</tr>
<tr>
<td>rs2155222</td>
<td>0.035</td>
<td>0.005 (0.025)</td>
<td></td>
</tr>
<tr>
<td>Whole blood</td>
<td>AP001189.4</td>
<td>Whole blood</td>
<td>Whole blood (GTEx Portal)</td>
</tr>
<tr>
<td>rs72944758</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>rs200331695 (IVS6-43A&gt;G)</td>
<td>*</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>rs10899221</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>rs12271056</td>
<td>*</td>
<td>9.8 x 10^-5</td>
<td></td>
</tr>
<tr>
<td>rs143975731</td>
<td>0.039</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>rs12277366</td>
<td>0.03</td>
<td>8.7 x 10^-5</td>
<td></td>
</tr>
<tr>
<td>rs2155225</td>
<td>*</td>
<td>1.5 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>rs2155222</td>
<td>*</td>
<td>1.4 x 10^-4</td>
<td></td>
</tr>
<tr>
<td>rs7126722</td>
<td>*</td>
<td>4.3 x 10^-4</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> The SNPs were analysed as a group due to correlation between variants; $r^2 = 1$.

<sup>*</sup> P or FDR > 0.05

Bold type signifies statistical significance.

Abbreviations: CRPC = castration-resistant prostate cancer, FDR = false discovery rate, GTEx Portal = Genotype-Tissue Expression Portal, NA = not analysed
Figure 7. The effects of prostate cancer-associated eQTLs on gene expression. The associations between genes and variants were identified in two independent sample sets (P or FDR < 0.05). The number of observations of each genotype is indicated in the round brackets below the genotypes. (A) The strongest effects of the variants on DGAT2 expression were observed in the replication set of tumours. Prostate cancer death-associated risk alleles of rs143975731 (G), rs12277366 (C), rs2155225 (T) and rs2155222 (A) decreased expression of this gene. (B) Prostate cancer death-associated allele of rs12277366 (C) increased the expression of AP001189.4 in the screening set of whole blood.

5.2.2.2 Haplotypes

Association analysis of haplotypes and gene expression was performed to further examine the regions influencing gene expression. A total of five haplotypes were found to affect gene expression including the intergenic prostate cancer death-associated SNPs. The SNPs rs12271056 and rs7126722 were included in three haplotypes with surrounding intergenic SNPs that affected the expression of *alkaline ceramidase 3* (*ACER3*) (FDR = 0.048). The SNPs rs143975731, rs12277366, rs2155225 and rs2155222 formed two haplotypes that were associated with the expression of *ACER3* (FDR = 0.034) or both *ACER3* and *p21 protein (Cdc42/Rac)-activated kinase 1* (*PAK1*) (FDR = 0.031 and 0.038, respectively). However, none of the aforementioned haplotypes were associated with prostate cancer or its clinicopathological features (FDR > 0.05); therefore, the possible contributions of alterations in the expression of these genes to prostate cancer susceptibility remains unclear.
6 Discussion

6.1 Contribution of genetic variation at 11q13.5 to prostate cancer susceptibility

6.1.1 Genetic variants and haplotypes

The screening and validation of both the target gene *EMSY* and the chromosomal region 11q13.5 resulted in identification of a rare risk variant (MAF < 1%) and common susceptibility variants (MAF > 5%). The common variants had weak to moderate effects on the risk of unselected prostate cancer (OR 1.1-1.3), which is generally similar to the effects of predisposition SNPs identified in GWASs (Deminichelis and Stanford, 2015). The associations of the intergenic SNPs with prostate cancer death were stronger than those to prostate cancer in general, indicating that these SNPs affect the life-threatening progression of this cancer. These findings may explain the lack of or weak associations with prostate cancer detected in the cases included in the screening arm of the PSA trial because PSA screening results in detection of non-significant cancers, in addition to clinically significant cancers (Draisma et al., 2003; Loeb et al., 2014). Similarly, the findings indicate that common *EMSY* variants increase the risk of initiation of cancer development because no associations with the clinicopathological features were observed and stronger associations were detected in the PSA screening cases. The rare *EMSY* SNP rs200331695 that increased the risk of familial prostate cancer, had a stronger effect on the disease risk (OR 7.5) compared to the common risk variants identified in this study. Overall, the findings of this thesis are consistent with the previously presented correlation between susceptibility variant allele frequency and penetrance (Figure 2) (McCarthy et al., 2008), and they indicate that both rare and common variants contribute to the development of prostate cancer, as suggested by the Common Disease, Common Variant (Reich and Lander, 2001; Schork et al., 2009) and Common Disease, Rare Variant hypotheses (Pritchard, 2001; Schork et al., 2009). These findings contribute to current knowledge regarding the genetic
susceptibility of prostate cancer and indicate that SNPs at 11q13.5 contribute to the development of this disease together with other susceptibility variants and somatic alterations.

A comparable finding of an inverse association of a prostate cancer risk SNP with advanced disease has been reported for a SNP located adjacent to a PSA-encoding gene \textit{kallikrein related peptidase 3 (KLK3)} (Kader et al., 2009; Pomerantz et al., 2011). The non-risk allele for prostate cancer is associated with a lower PSA level in controls; thus, hypothetically, men carrying this allele are examined for prostate cancer less frequently, which may lead to development of advanced disease, i.e., the observed association with advanced disease is due to a PSA bias (Kader et al., 2009; Pomerantz et al., 2011). Here, we were not able to test the associations of the SNPs with the PSA level in controls, but no associations were observed in the prostate cancer patients. The effects of the SNPs on \textit{KLK3} expression were not analysed in our sample sets because we only studied \textit{cis}-effects of the variants, but no associations between SNPs (rs12271056, rs12277366, rs2155225, rs2155222 and rs7126722) and \textit{KLK3} expression in prostate tissues were detected using GTEx Portal v6 (GTEx Consortium, 2015; GTEx Web Portal). Therefore, it is unlikely that a PSA bias affected the results here. Interestingly, according to recent analysis, the detection of some of the GWAS-identified risk SNPs may have been a result of a PSA bias, if SNPs were associated with PSA concentration, and this issue should be considered when interpreting results from a GWAS (Dluzniewski et al., 2015).

These data indicate that at least two independent association signals are present in 11q13.5, as the \textit{EMSY} SNPs were separated from the high-LD intergenic SNPs by a recombination hotspot (Kong et al., 2002; Kong et al., 2010; UCSC Genome Browser), and variants of \textit{EMSY} and the intergenic region were located in different haploblocks. In addition, no LD was observed between the \textit{EMSY} variants and intergenic SNPs, and only the intergenic SNPs were found to be associated with prostate cancer death. The independence of the association signal of the \textit{EMSY} SNP rs200331695 in relative to those of the other SNPs is not clear because LD between rare and common variants is difficult to analyse, and haploblock analyses included only common variants (\textbf{I} and \textbf{II}) or common haplotypes (\textbf{III}).

The risk SNPs identified through GWAS only explain a portion of the genetic heritability of prostate cancer and other complex diseases (Kim et al., 2010; So et al., 2011). The remaining heritability has been suggested to be attributed to, for example, low-frequency and rare variants, structural variations, genetic interactions between loci (epistasis), gene-environment interactions and epigenetic effects (Eichler et al., 2010; Manolio et al., 2009; Zuk et al., 2012). Interestingly, haplotypes composed of
common SNPs have been suggested to account for the remaining heritability (Bhatia et al., 2015). Here, a total of four haplotypes were identified in 11q13.5 that contributed to prostate cancer susceptibility. The EMSY haplotypes could explain the associations observed between the high-frequency EMSY alleles and prostate cancer predisposition because the frequencies of these haplotypes were lower. In addition, inclusion of the risk alleles of EMSY SNPs in a both predisposing and protective haplotype could indicate involvement of genetic interactions between SNPs, which would modulate the effect. However, because the haplotypes were composed partly of imputed SNPs and were defined using an in silico method, the haplotype findings require further experimental validation.

Most cancer-related studies of EMSY have focused on copy number alterations and the effects of overexpression of this gene in cells; however, to the best of our knowledge, no studies have investigated its association with prostate cancer. Therefore, the results of this thesis are novel and contribute to current knowledge regarding EMSY as a cancer-related candidate gene. Associations of EMSY variants with breast and ovarian cancers have been previously evaluated, but no predisposing variants have been detected (Benusiglio et al., 2005), in contrast with our findings. However, the breast and ovarian cancer study used a tag SNP approach to detected variants in this gene, in addition to screening of the promoter and coding regions in a fairly small number of breast cancer cases (n = 48); thus, disease-predisposing SNPs, especially rare variants, might have been missed (Benusiglio et al., 2005).

Although prostate cancer and aggressive prostate cancer risk-increasing variants in EMSY were observed in this thesis, the role of EMSY as a prostate cancer candidate gene remains unclear because effects of the intronic variants on the splicing of EMSY messenger RNA (mRNA) were only predicted by in silico analysis, and EMSY was not found to be regulated by the prostate cancer risk SNPs in whole blood, LCLs, or prostate tumours. Interestingly, associations of the common intronic SNPs rs72944758 and rs10899221 with expression of EMSY in normal prostate tissues were identified using the GTEx Portal, suggesting a possible role of EMSY alteration during the initiation phase of cancer development, which is supported by the association findings for the SNPs. The results regarding the roles of EMSY variants as EMSY regulators in prostate cancer require further validation.
6.1.2 Prostate cancer death-associated risk SNPs affect \textit{AP001189.4} and \textit{DGAT2} expression

All identified prostate cancer-associated variants were located within non-coding areas, either in introns or in intergenic regions, which directed the functionality studies of the effects of these variants on gene expression. Overall, current research on prostate cancer susceptibility variants has focused more on understanding how the risk SNPs contribute to the disease rather than merely on the identification of new susceptibility variants. One of the most frequently applied approaches to studying regulatory variants is eQTL mapping (Albert and Kruglyak, 2015), which was also used in this thesis.

Functional annotations were available for two of the tissue types in which the regulatory effects of the variants were analysed: the peripheral blood and LCLs. The rare \textit{EMSY} SNP rs200331695 was observed to affect the expression of \textit{DGAT2} in LCLs, but the functional annotation did not support this finding. However, the data on functional elements for the LCLs were based only on a single cell line of the ENCODE project (ENCODE Project Consortium et al., 2012). Overall, this association requires further validation using an independent set of LCLs because it could not be replicated using the GTEx Portal due to a missing variant. The prostate cancer death-associated intergenic SNPs coincided with enhancer elements in one or several cell types in peripheral blood (T regulatory, B, and natural killer cells), in support of the replicated finding that the intergenic SNP rs12277366 is an eQTL of \textit{AP001189.4} in whole blood. \textit{AP001189.4} encodes a long non-coding RNA (lncRNA) of unknown function (Ensembl genome browser; Flicek et al., 2014). No reports have been published on \textit{AP001189.4}, as determined based on a PubMed search (PubMed), and this gene has not been deposited in Long Noncoding RNA Database v2.0 (Amaral et al., 2011; lncR\textsc{NAdb}; Quek et al., 2015), highlighting the novelty of this finding of a disease-associated gene. Non-coding RNAs, which have been previously considered to be merely junk RNA, are at least partially functional (Palazzo and Lee, 2015) and affect various processes that are important for tumour development, including proliferation, metastasis, self-renewal, survival and apoptosis, through transcriptional or post-transcriptional regulation (Su et al., 2015). At least 13 well-studied lncRNAs have been associated with prostate cancer, and some of the identified lncRNAs have been suggested to be biomarkers for prostate cancer (Su et al., 2015). As \textit{AP001189.4} was identified in whole blood and the
functional annotation was related to cells of the immune system, its possible mode of action could involve immune system processes, which are connected with cancer development (de Visser et al., 2006; Hanahan and Weinberg, 2011). Interestingly, prostate tumour stroma, which includes immune cells, has been reported to contribute to malignant transformation of the prostate and to co-develop with tumour cells, for example, by altering expression levels (Dakhova et al., 2009; Planche et al., 2011). Expression alterations detected in stroma have been suggested to be potential biomarkers that could be used in prostate cancer diagnosis (Zhu et al., 2015).

Another replicated target gene of prostate cancer death-associated SNPs was DGAT2 in prostate tumours. Unfortunately, no functional annotations were available for prostate tumours; therefore, the existence of regulatory motifs specific to prostate tumours in relation to the identified eQTLs remains unclear. The SNPs were more strongly associated with DGAT2 expression in the validation set of tumours, which consisted solely of specimens from untreated prostatectomies, compared to the screening set. The inclusion of CRPCs in the screening set might explain this difference, for example, because of treatment before sampling. However, due to missing patient information for most of the samples (7/11), this issue could not be further analysed. Another difference between the sample sets was the time of onset of biochemical recurrence (increase in the PSA level) after prostatectomy. The median time for the screening set was 2.1 years, which can be classified as early onset (Shahabi et al., 2016), whereas that for in the validation set was 4.2 years, which can be classified as late onset (Shahabi et al., 2016). Some factors other than the SNPs might have influenced the expression of DGAT2, as its expression has been reported to be associated with the biochemical recurrence of prostate cancer (Heemers et al., 2011), whereas no associations were detected between the SNPs and an increased PSA level after treatment.

The prostate cancer death-associated alleles were observed to contribute to decreased DGAT2 expression. DGAT2 catalyses the biosynthesis of triacylglycerol lipids (Cases et al., 2001); thus, hypothetically, a low level of DGAT2 might result in low levels of triacylglycerols in prostate tumours. Low levels of these lipids are present at least in colorectal tumours, especially in patients with lymph node metastasis; however, the mechanism contributing to this phenomenon is unknown (Zhang et al., 2014). According to analysis performed using the GTEx Portal data, the SNPs did not have regulatory effects on DGAT2 in the normal prostate tissues, in support of the hypothesis that SNPs contribute to tumour development through DGAT2 during a later stage of tumourigenesis. Androgens have been reported to
downregulate the expression of DGAT2 in prostate cancer cell lines (Bu et al., 2016; Heemers et al., 2011; Ngan et al., 2009; Rajan et al., 2011), similar to our findings of the risk SNPs in tumours. Prostate cancer development is androgen dependent (Andriole et al., 2010; Thompson et al., 2003), and after androgen deprivation therapy, the disease may develop into life-threatening CRPC with reactivated androgen receptor (AR) signalling (Yuan et al., 2014). Based on the functional annotations of the prostate cancer death-associated SNPs, rs143975731 affects the motif of transcription factor upstream binding protein 1 (LBP-1a) (UBP1, alias LBP1), which has been suggested to act as an AR co-regulator (Lin et al., 2009). Thus, the risk SNPs studied here may have androgen-dependent effects on DGAT2 expression. Interestingly, a recent study has reported that many prostate cancer risk-associated loci influence binding of the AR-forkhead box A1 (FOXA1) or AR-HOXB13 complexes and thereby function in gene regulation (Whitington et al., 2016).

The finding of tissue-specific regulation of two genes by the same high-LD SNPs is interesting because association analysis revealed that different alleles of SNPs increased the risk of either prostate cancer in general or prostate cancer death. The two identified target genes may increase the risks of different types of prostate cancer or function during different phases of tumour development in different tissues, which would explain the observed associations. Overall, our eQTL mapping results are in agreement with previous observations that non-coding risk variants act as eQTLs and contribute to prostate cancer susceptibility, possibly through gene regulation (Amin Al Olama et al., 2015; Chen et al., 2015; Han et al., 2015b; Jiang et al., 2014; Li et al., 2014; Xu et al., 2014).

6.1.3 Prostate cancer-associated chromosomal region 11q13-14

Both linkage and association studies have emphasized the importance of the chromosomal region 11q13-14 in prostate cancer susceptibility. Linkage within 11q14.1 was specifically detected in Finnish prostate cancer families (Schleutker et al., 2003), which were also screened for EMSY variants. However, none of the 11q14.1-linked families carried the rare EMSY mutation rs200331695, which was observed to increase the risk of aggressive cancer. The associations of the other identified prostate cancer risk SNPs with familial prostate cancer were not analysed, but it is unlikely that they would, at least individually, explain the linkage signal, as they are common, low-risk variants, and linkage analyses are more suitable for
identifying rare, high-risk variants. However, it is possible that the identified variants in *EMSY* and in the intergenic region have multiplicative effects on the disease risk. These effects should be evaluated in the 11q14-linked families; however, it is possible that an additional susceptibility factor is present in the prostate cancer-linked region and it remains to be identified.

The prostate cancer-associated chromosomal region 11q13.3, which was originally identified by two GWASs (Eeles et al., 2008; Thomas et al., 2008), is located approximately 7 Mb from the chromosomal region 11q13.5 identified in this thesis. Similarities exist between these two regions because the association signals originate from non-coding SNPs, and they both appear to contain more than one independent associated region (Amin Al Olama et al., 2015; Chung et al., 2011; Zheng et al., 2009). In addition, a potential AR-mediated mechanism has been suggested for 11q13.3 because rs11351679, which is a proxy LD SNP of the lead SNP of the locus, disrupts binding of the AR-FOXA1 complex (Whittington et al., 2016). However, several factors indicate that the associations observed in 11q13.5 are novel and independent of associations in 11q13.3. First, multiple recombination hotspots separate these regions from each other (Kong et al., 2002; Kong et al., 2010; UCSC Genome Browser). Second, the SNPs in 11q13.5 are not in LD ($r^2 > 0.2$) with those in 11q13.3 [LD analysis conducted using Haploreg v4.1 (Ward and Kellis, 2012)]. Third, the 11q13.5 SNPs were found to be associated with prostate cancer death but no such associations have been observed for 11q13.3 SNPs (Penney et al., 2010). Interestingly, loci in 11q12.2 and 11q14.1 have been suggested to be associated with prostate cancer mortality, although genome-wide significances were not observed for these findings (Penney et al., 2010). No LD ($r^2 > 0.2$) exists between the SNPs within the aforementioned regions and those within 11q13.5, indicating that the signals are independent [LD analysis conducted using Haploreg v4.1 (Ward and Kellis, 2012)]. Although no connection appears to exist between the 11q13.5 and 11q13.3 SNPs, the existence of shared regulatory target genes cannot be excluded based on the available data because eQTL mapping studies have frequently analysed regulatory effects *in cis*, including only genes within a 1-Mb region surrounding the variants. *Trans*-effects of an 11q13.3 SNP has been analysed in tumour stroma only (Chen et al., 2015), but no data of stroma are available concerning the 11q13.5 SNPs identified in this thesis. However, the *DGAT2* and *AP001189.4* genes observed here were not identified as target genes of the 11q13.3 eQTL in tumour stroma (Chen et al., 2015).
6.2 Methodological considerations

The variants were screened in this thesis using Sanger sequencing and imputation, which differ from each other in many aspects. The fundamental difference between these two methods is that Sanger sequencing generates genomic sequences with high accuracy, and imputation computationally calculates genotypes. The accuracy of imputation depends on multiple factors, such as the MAFs of untyped SNPs, LD level, marker density and reference population (Pei et al., 2008). Here, the imputed genotypes were confirmed by genotyping, and all associations were validated. Sanger sequencing is suitable for detecting all variations within a target region in contrast with imputation, which defines genotypes of known variants described in a reference population but misses undescribed variations. For example, the screening of EMSY using Sanger sequencing resulted in detection of a total of 17 variants (63 % of all identified variants) that were not present in Single Nucleotide Polymorphism Database (dbSNP) (dbSNP; Sherry et al., 2001) at the time of publication. The disadvantage of Sanger sequencing is its laboriousness. It is time-consuming and the resources limit the length of the target region that can be analysed. In contrast, imputation enables coverage of larger genomic regions at the expense of computational burden. The fine mapping of 11q13.5 demonstrated the disadvantages of Sanger sequencing because two prostate cancer risk SNPs located in deep intronic regions of EMSY were not identified by sequencing but were detected by the imputation method. In summary, the fine mapping of 11q13.5 performed in this thesis resulted in good coverage of the known common variants within this region but missed the undescribed variants, which were not located within the regions of EMSY that were screened in detail. In addition, possible prostate cancer-associated low-frequency and rare known variants were probably missed because such variants are challenging to impute (Zheng et al., 2015).

The genotyping of variants was performed using TaqMan® (Applied Biosystems) and KASPar™ assays (KBioscience), as well as high-resolution melt analysis (HRMA) and Sanger sequencing. The accuracies of both TaqMan® and KASPar™ assays are high according to the manufacturers, although they use different chemistries for mutation detection. TaqMan® assay involves the use of allele-specific labelled hybridization probes, whereas KASPar™ assay involves the use of allele-specific primers with unique tails and labelled universal nucleotide cassettes that are complementary to primer tails. Because no labelled probes are used in KASPar™ assay, the cost is reduced. However, genotyping assays, which are frequently performed by researchers, are not available for all SNPs, necessitating the
use of alternative methods, such as Sanger sequencing or HRMA. HRMA can be used to both screen and genotype variants, and it has been reported to be a powerful technique for detecting human disease-associated mutations with high accuracy, low cost and simplicity (Li et al., 2011).

The advent of next-generation molecular technologies has enabled the generation of massive amounts of data, for example, whole-genome or transcriptome data, in a short period of time, and the available new techniques, such as NGS and next-generation genotyping platforms, have at least partly replaced the traditional methods used for variant detection and genotyping as well as determination of gene expression levels. RNA sequencing was performed in this thesis to measure gene expressions during the screening phase of eQTL mapping. RNA sequencing has been established as an accurate method for measuring relative gene expression (SEQC/MAQC-III Consortium, 2014). However, massive amounts of data are produced, which are difficult to handle and analyse (Han et al., 2015a), thereby affecting the level of confidence and reproducibility of the results. Thus, the preliminary eQTL analysis results were validated using traditional RT-qPCR to measure gene expression, and the association of one out of four genes was replicated.

As new techniques are developed, the amounts of both genetic variation and functional data increase and become available in public databases. Thus, the functional annotations of the prostate cancer-associated variants were reanalysed using updated versions of HaploReg v4.1 (Ward and Kellis, 2012) and HumanSplicingFinder v3.0 (Desmet et al., 2009) in this thesis. Specifically, the latest version of HaploReg (v4.1) included integrated data from the Roadmap Epigenome Project and ENCODE Project, which is the most comprehensive map of the human genomic landscape created for use in studies of, for example, genome interpretation, gene regulation, genetic variation and human disease (Roadmap Epigenomics Consortium et al., 2015). Briefly, data on the functional elements and open chromatin coinciding with the prostate cancer risk SNPs differed between original analysis and reanalysis because the previous version of HaploReg (v2) included provisional Roadmap data in addition to data from ENCODE, and these two data sets were not integrated. In contrast, the annotation results regarding protein binding and predicted changes to the regulatory motifs were similar between the analyses. However, the conclusions of both analyses conducted with different versions of HaploReg were similar; functional annotations indicated functionality of the variants. The greatest difference between analyses was observed for the rare EMSY SNP rs200331695; original analysis found no functionality of this SNP, whereas reanalysis identified possible effects on splicing and regulatory motifs. These
discrepancies could merely be due to differences between software algorithms because different web tools were used in the two analyses. Large updates were not included in the latest version of Human Splicing Finder v3.0 (Desmet et al., 2009) because the results were similar for the two other intronic EMSY SNPs, which were originally analysed with the same software.

6.3 Samples

Evaluation of a population of individuals with mixed ancestry, i.e., with population stratification, is a challenge, particularly in association studies, due to varying allele frequencies irrespective of the disease statuses of individuals with different genetic backgrounds (Cardon and Palmer, 2003). In this study, all samples were obtained from individuals from the Finnish population, which is genetically homogenous due to its demographic history (Peltonen et al., 2000) and is therefore considered suitable for genetic studies of complex traits (Kristiansson et al., 2008; Peltonen et al., 2000). However, genetic differences have been found to exist among different parts of Finland, which might affect the results of association studies (Palo et al., 2009; Salmela et al., 2008).

The availability of appropriate controls for the study of late-onset disease is limited; thus, healthy Finnish blood donors were used as controls. However, the use of blood donors is not ideal because they are younger (age of 18-65 years) (Finnish Red Cross Blood Service) than the average age of prostate cancer onset (approximately 70 years) (Pukkala et al., 2011). The use of such controls may have decreased the strength of the detected associations, as previously described for a HOXB13 mutation in the Finnish population (Laitinen et al., 2013), due to the presence of undetected and developing tumours in individuals misclassified as controls. In addition to population controls, prostate cancer families include unaffected individuals who have not yet reached the age of onset, which hinders family-based analyses. For this reason, the information on the individuals in the rs200331695-positive families was updated. Few unaffected mutation-positive family members had reached the average age of onset at the time of this update but no additional prostate cancer diagnoses had emerged since publication of the results of study I. However, samples from two nonaggressive cases had become available for genotyping, and the mutation-negative statuses of the aforementioned patients in family 017 further confirmed the finding of occurrence of rs200331695 with an aggressive disease outcome.
Due to PSA testing, prostate cancer cases are enriched with early-stage prostate cancer, which adds complexity to the study of this genetically heterogeneous disease, and it has been suggested that studies should focus on specific disease subtypes (Platz et al., 2004). However, no standard definitions of the clinically significant or aggressive form of prostate cancer exist. Notably, the definition of aggressive prostate cancer differed between studies I and II. The criteria used in study I were modified from the ICPCG criteria (Schaid et al., 2006) and aggressiveness was determined based on the perspectives of urologists in study II (personal communication). Therefore, the findings of these two studies are not completely comparable. Overall, to improve the study of susceptibility variants and the reproducibility of findings, more precise definitions should be established to avoid the misclassification of study subjects. Such efforts have been recently put forth by the International Society of Urologic Pathology, which introduced new risk categories based on a five-grade Gleason grading system (Epstein et al., 2016). Interestingly, the frequency of genetic aberrations has been found to be correlated with the proposed risk grades (Rubin et al., 2015). These new risk categories have been accepted for inclusion in the new 2016 World Health Organization guidelines (Epstein et al., 2016).

Due to the tissue specificity of trait-associated eQTLs (Fu et al., 2012), the effects of the identified prostate cancer-associated variants were analysed in multiple tissues. Previous findings have suggested that LCLs are good models for use in gene expression studies (Bullaughhey et al., 2009; Ding et al., 2010; Nicolae et al., 2010); thus, the LCLs were chosen for validation of the eQTL findings for whole blood in this thesis. However, our results suggested that there were differences in gene expression regulation between LCLs and other tissues, which has been reported in other studies (Powell et al., 2012; Sasayama et al., 2013; Xu et al., 2014). A subset of variants were identified that particularly increase the risk of prostate cancer death. Thus, it would have been ideal to study their effects in more advanced tumour types; however, such an analysis is limited by the availability of tumour samples. In addition, the accessibility to tumour specimens in general affects the sample size on which the results are based. According to the GTEx Portal, a minimum of 70 samples per tissue type are required to identify a cis-eQTL (GTEx Web Portal). Here, the eQTLs targeting AP001189.4 were identified in sets of 99 and 338 whole blood samples, and the results for the DGAT2 eQTLs were based on a total of 82 tumour samples, which exceeded the limit suggested by the GTEx Portal.
6.4 Future prospects

This study has revealed interesting and novel findings regarding genetic susceptibility to advanced prostate cancer; however, further study is necessary, particularly to replicate the associations of the SNPs in other populations and to evaluate the functional consequences of the SNPs. A causal variant, a regulatory motif that affects expression of the identified genes, and transcription factors that modulate these effects should be identified. In addition, the effects of the SNPs on gene expression regulation must be confirmed at the protein level for DGAT2, and the downstream effects of alterations of both AP001189.4 and DGAT2 in cells need to be investigated. Because the functional annotations indicated that intronic EMSY SNPs affect splicing, splicing analysis would be of interest, especially because a phosphorylation site regulating the transcriptional activity of EMSY (Ezell et al., 2012) is encoded by an exon surrounded by the introns in which the variants are situated. In addition, the effects of the variants on gene expression in trans would be interesting to study, in addition to their effects on gene expression in other tissues, such as normal prostate tissues and tumour stroma. Furthermore, replication of the findings of the regulatory effects of the SNPs using samples from other populations could be conducted, as no major differences in cis-eQTLs have been found to exist among ethnic groups (Stranger et al., 2012).

The findings of this thesis contribute to current knowledge regarding genetic susceptibility to prostate cancer, and the identified genetic variants have a potential to be used as biomarkers in the future. The aggressive and prostate cancer death-associated SNPs could be utilized in determining the risk of advanced cancer in men at screening to assess the need for further examination and biopsies. The rare EMSY-variant in particular could be used in family-based screening of prostate cancer. Further, these variants could be used to identify prostate cancer patients at high risk of life-threatening disease progression, which would enable proper treatment of these patients.

In addition to risk management, disease-associated germline variants have a great potential for use in the development of personalized medicine. In personalized medicine, patients are treated based on their genetic backgrounds or other specific features characteristics of particular diseases to maximize treatment benefits (Juran and Lazaridis, 2011). For example, targeted therapy based on poly-ADP-ribose polymerase (PARP) inhibitors has been recently approved for treatment of ovarian cancer patients with germline BRCA mutations (Liu and Matulonis, 2016). The same therapy has been tested in breast and prostate cancer patients with BRCA mutations,
and the results have been promising (Livraghi and Garber, 2015; Mateo et al., 2015). However, further functional studies regarding the findings of this thesis must be conducted before such therapies could be developed.

A joint analysis of the risk SNPs identified in this thesis and in others studies would be an ideal approach to establish a multiSNP risk profile that could be used to further assess the potentials of these SNPs as biomarkers. For example, they could be analysed together with other risk SNPs that are detected in Finns, such as variants in chromosomes 2q and 17q including \textit{HOXB13} (Laitinen et al., 2013; Laitinen et al., 2015), in \textit{interleukin 10 (IL10)} (Faupel-Badger et al., 2008), in \textit{ADP-ribosylation factor like GTPase 11 (ARL11)} (Siltanen et al., 2011) and in \textit{checkpoint kinase 2 (CHEK2)} (Seppala et al., 2003). Multiple prostate cancer-specific multiSNP panels have been evaluated for their ability to be used in the detection of prostate cancer, predicting of risk for this cancer or to be used as prognostic indicators in affected men but no currently available SNP panel is clinically useful for predicting the risk of prostate cancer, for distinguishing between aggressive and asymptomatic disease or for predicting prostate cancer mortality (Little et al., 2015). Identification and validation of additional risk variants, in particular causal variants, would enable larger SNP panels to be developed that could offer greater clinical significance than the current SNP panels (Little et al., 2015). In addition, combination of different types of biomarkers could improve their potential to be used in risk prediction, in screening and in diagnostics. For example, a Stockholm 3 (STHLM3) model, which combines information of six plasma protein biomarkers, over 200 prostate cancer-associated SNPs and clinical data (age, family history, prostate biopsies and examinations), has shown promising results in men 50-69 years old in detecting aggressive cancer and reducing the number of unnecessary biopsies in those with clinically insignificant cancer (Gronberg et al., 2015).

Obviously, further research is required before discoveries of genetic research are applied in routine clinical practice. In addition, mathematical, statistical and bioinformatics expertise will be required to improve the interpretation of cancer risk (Thomas et al., 2015). Future cancer research will exploit various types of genomic data using next-generation technologies and conduct integrative network analyses to obtain a more detailed view of the origins of cancer (Tang et al., 2013); however, more collaboration between basic genomics researchers and clinical researchers is needed before discoveries can be translated into the clinical settings (Thomas et al., 2015). The emerging biobanks will permit the widespread use of samples, and new collaborative projects will enable both data and knowledge to be shared in favour of research.
7 Summary and Conclusions

The comprehensive collection of samples and patient information from both unselected prostate cancer cases and prostate cancer families by the University of Tampere and the Tampere University Hospital enabled the study of genetic risk factors for prostate cancer. This thesis targeted the search for risk variants to the chromosomal region 11q13.5 and aimed to specifically identify genetic variants that predispose men to the advanced form of prostate cancer. This thesis utilized different types of samples and patient information, as well as new and traditional research techniques, to obtain interesting and novel findings regarding genetic susceptibility to prostate cancer. The major conclusions are as follows:

1. The chromosomal region 11q13.5 is associated with prostate cancer risk, with a complex genetic structure, including both rare and common SNPs, in addition to haplotypes.
2. The genetic variants within 11q13.5 increase susceptibility to both prostate cancer in general and the advanced form of this disease.
3. The intronic risk SNPs in EMSY possess functional potentials, and may result in alterations in the EMSY protein.
4. The prostate cancer death-predisposing intergenic SNPs affect the expression of DGAT2 in prostate tumours and that of AP001189.4 in whole blood, suggesting possible contributions of these SNPs to prostate cancer development through tissue-specific gene regulation.
Acknowledgements

This study was performed at the Laboratory of Cancer Genetics, BioMediTech, University of Tampere, Tampere, from 2009-2016. I want to thank the former director of IBT, Professor Olli Silvennoinen, M.D., Ph.D., the current director of BioMediTech, Dr. Hannu Hanhijärvi, DDS, Ph.D., and the medical director of the Fimlab Laboratories, Docent Erkki Seppälä, M.D., Ph.D., for providing excellent research facilities for carrying out the thesis project. In addition, I wish to acknowledge the Tampere Graduate Program in Biomedicine and Biotechnology (TGPBB) for organizing interesting courses and for enabling me to participate in international research meetings by providing travel grants. I thank the following parties for providing funding for this research: the Competitive Research Funding of the Tampere University Hospital, the Finnish Cancer Organisations, the Reino Lahtikari Foundation, the Sigrid Juselius Foundation, the Academy of Finland, the Cancer Society of Finland, the Medical Research Fund of Tampere University Hospital, the Worldwide Cancer Research, the Finnish Cultural Foundation, the Emil Aaltonen Foundation, the Ida Montin Foundation, the Scientific Foundation of City of Tampere and the Maud Kuistila Memorial Foundation.

I sincerely thank all of the cancer patients and their family members who volunteered to participate in this study.

I wish to thank all the people who have contributed to the thesis process. First, I sincerely thank my supervisors Professor Johanna Schleutker, Ph.D., and Docent Tiina Wahlfors, Ph.D., for their expertise, guidance and support. Johanna gave me the opportunity to work in her research group and to begin my scientific career in the fascinating field of cancer genetics, which I greatly appreciate. This opportunity has taught me a lot about science and being a researcher. Tiina, our post-doc, guided me through many practical issues during the thesis process, and her enthusiasm towards scientific research was very inspiring.

I also wish to thank my thesis committee members, Professor Markku Kulomaa, Ph.D., and Docent Janna Saarela, M.D., Ph.D., for their feedback on this thesis project.
I additionally thank the official reviewers of the thesis manuscript, Docent Pia Vahteristo, Ph.D., and Docent Peter Boström, M.D., Ph.D., for their valuable comments and constructive criticism.

I sincerely thank all of the co-authors for their professional help and contribution to this work. In particular, I am truly thankful to Professor Teuvo Tammela, M.D., Ph.D., Professor Tapio Visakorpi, M.D., Ph.D., and Professor Anssi Auvinen, M.D., Ph.D., for granting me with access to the comprehensive patient material on which this thesis was based. I am also very grateful for all of the assistance that I received with the statistical aspects. I greatly acknowledge the contributions of Tommi Rantapero, M.Sc., Daniel Fischer, M.Sc., Docent Rainer Lehtonen, Ph.D., Swee Wong, M.Sc., and Professor Matti Nykter, Ph.D.

I was lucky to have the opportunity to work with the skilful members of the Genetic Predisposition to Cancer study group. In particular, I am grateful to Henna Mattila, Ph.D., Sanna Siltanen, Ph.D., Kirsi Määttä, Ph.D., Virpi Laitinen, M.Sc., CMG, Ms. Riina Kylätie and Ms. Linda Enroth, for all of the memorable moments in and out of work and for helping to create a friendly working atmosphere. I would also like to thank Ms. Riitta Vaalavuo for her valuable assistance with the patient information, Elisa Vuorinen, M.Sc., and Ekaterina Slitikova, M.Sc., for all of the help in the laboratory and all of the group members who I worked with for shorter periods of time. In addition, I thank “the coffee room people” for the daily chit chat.

I am deeply grateful for all of the support from friends and family. I thank Jenni, Jonna, Minttu, Niina and Elisa, with whom I share a common interest in biology and genetics, for all of the peer support. In addition, I thank Jenni, Maija, Minna and other friends for all of the get-togethers, which were truly needed to keep the balance. With warm thoughts of gratitude and appreciation, I thank my sisters and most of all my friends, Kirsi and Minttu, as well as my parents, Mirja and Pekka, for being there for me. Most importantly, to the men in my life, Joonas and Topi: ♥

Tampere, May 2016

Riikka Nurminen
References


Cousineau, I., and Belmaaza, A. (2011). EMSY overexpression disrupts the BRCA2/RAD51 pathway in the DNA-damage response: implications for chromosomal


Original Communications
Identification of an aggressive prostate cancer predisposing variant at 11q13

Riikka Nurminen¹, Tiina Wahlfors¹, Teuvo L. J. Tammela² and Johanna Schleutker¹

¹ Laboratory of Cancer Genetics, Institute of Medical Technology and Centre of Laboratory Medicine, University of Tampere and Tampere University Hospital, Biokatu 8, FI-33014 Tampere, Finland
² Division of Urology, Tampere University Hospital and Medical School, Teiskontie 35, PO BOX 2000, FI-33521 Tampere, Finland

Prostate cancer is the most frequently diagnosed cancer in men; however, the genetic basis of susceptibility remains elusive. The EMSY gene is located in the prostate cancer linked chromosome region at 11q13.5. The aim of this study was to screen EMSY for sequence variants and to evaluate its association with the risk of prostate cancer. We performed a Finnish population-based case–control study with 923 controls, 184 familial prostate cancer cases and 2,301 unselected prostate cancer cases. Variants were screened using sequencing and validated using the TaqMan assay and High Resolution Melting analysis. A total of 27 sequence variants were found, and 17 of them were novel. A rare intronic variant, IVS6-43A>G (minor allele frequency of 0.004), increased the prostate cancer risk in familial cases (odds ratio [OR] = 7.5; 95% confidence interval [CI] = 1.3–45.5; p = 0.02). Further analysis with clinicopathological data revealed that the variant is associated with aggressive unselected cases (prostate specific antigen ≥ 20 μg/L or Gleason grade ≥ 7), based on both case–control (OR = 6.0; 95% CI = 1.3–26.4; p = 0.03) and case–case analyses (OR = 6.5; 95% CI = 1.5–28.4; p = 0.002). In addition, all variant-positive familial cases had aggressive cancer. Our results indicate that the intronic variant IVS6-43A>G increases the familial and unselected prostate cancer risk in a Finnish population and contributes to the aggressive progression of the disease in a high-penetrance manner. The potential role of the variant as a predictive genetic marker for aggressive prostate cancer should be further evaluated.

Introduction

In Finland, as in many western countries, prostate cancer (PrCa) is the most frequently diagnosed cancer and the second most frequent cause of cancer related deaths in men.¹

Key words: aggressiveness, association, EMSY, prostate cancer genetics, 11q13

Abbreviations: CI: confidence interval; ENT domain: EMSY N-terminal domain; GWAS: genome-wide association study; HRM: high resolution melt; HWE: Hardy-Weinberg equilibrium; ICPGC: International Consortium for Prostate Cancer Genetics; LD: linkage disequilibrium; MAF: minor allele frequency; OR: odds ratio; PrCa: prostate cancer; PSA: prostate specific antigen; SNP: single nucleotide polymorphism

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Competitive Research Funding of the Tampere University Hospital; Grant number: 9L091; Grant sponsors: the Reino Lahtikari Foundation, the Finnish Cancer Organizations, the Sigrid Juselius Foundation, the Academy of Finland; Grant numbers: 126741, 116437

DOI: 10.1002/ijc.25754

History: Received 29 Apr 2010; Accepted 30 Sep 2010; Online 25 Oct 2010

Correspondence to: Johanna Schleutker, Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, Biokatu 8, FI-33014 Tampere, Finland, Fax: +358-3-3117-4168, E-mail: Johanna.Schleutker@uta.fi

The majority of cases are sporadic (75–80%),² while approximately 5–10% of PrCa cases result from heritable factors.³ According to a Nordic twin study, the contribution of heritable factors to the risk of PrCa could be as high as 42%,⁴ which is the highest ever reported for a common malignancy. A variety of inheritance patterns for familial PrCa have been proposed, including autosomal dominant,⁵ autosomal recessive and X-chromosomal inheritances,⁶,⁷ which reflects the very heterogeneous nature of the disease.

We previously reported two new Finnish hereditary PrCa linked loci at 11q and 3p.⁸ The most significant linkage signal for chromosome 11 was detected between the D11S1314 and D11S898 (11q13.4–11q22) markers, with the highest peak at D11S901 (11q14). Connections between PrCa and 11q have also been observed in genome-wide association studies (GWAS), which reported PrCa associated single nucleotide polymorphisms (SNPs) at 11q13.2 (rs7931342⁹ and rs10896449¹⁰). The SNP rs7931342 is located in a gene desert region,⁹ and rs10896449 is 67 kb upstream of the MYELOV [myeloma overexpressed (in a subset of t(11;14) positive multiple myelomas)] gene.¹⁰ The signal on 11q13 has been confirmed by other groups¹¹–¹³ and refined by a PrCa genome-wide association follow-up study.¹⁴

The importance of the region is further supported by a fine mapping study, in which two independent loci (rs10896449 and rs12418451) at 11q13 were associated with PrCa risk.¹⁵ In addition to the associated SNPs, DNA copy number changes have been observed at 11q13–11q14,¹⁶–¹⁹

but a candidate gene that predisposes patients to PrCa has not been identified.

The increasing number of new PrCa cases creates pressure on the health care system. For the development of new and improved tools for PrCa prevention, diagnosis, prognosis and treatment, it is essential to distinguish between men who are at risk of aggressive PrCa from those at risk for the majority of cancers that never develop into the life-threatening stage. Despite attempts to find predisposing variants, it still remains uncertain whether a genetic basis for the development of aggressive PrCa exists. One of the most recent association studies supports the idea that there are aggressive PrCa predisposing variants in the genome.20 In addition, epigenetic variants have been reported to be involved in the aggressive progression of PrCa,21 but these alternations are not necessarily involved in predisposition.

A new Breast cancer gene 2 (BRCA2) interacting protein EMSY was identified in 2003.22 The EMSY gene has been mapped to the chromosome locus 11q13.4–13.5.22 Amplification of the gene has been studied in many cancer types, such as breast, ovarian and pancreatic cancers,23–26 but its role in PrCa has not yet been evaluated. EMSY encodes a 1322 amino acid protein that localizes exclusively to the nucleus.22 At the amino terminus of EMSY, there is an evolutionarily conserved EMSY N-terminal (ENT) domain, which is unique in the human genome.22 The N-terminal part of the protein is also involved in the interaction with BRCA2,22 a protein that functions in recombinational DNA repair27 and transcriptional regulation.28 EMSY has been reported to bind to the transactivation domain of BRCA2, which consequently inhibits the activity of BRCA2, and to localize to DNA damage sites, which further supports its function in DNA repair.22 In addition to BRCA2, EMSY interacts with the chromatin remodeling proteins Heterochromatin protein 1 (HP1) and BS69.22 EMSY also forms a complex with NRC interacting factor 1 (NIF1) and deleted in breast cancer 1 (DBC1), and this protein complex plays an important role in the regulation of nuclear receptor-mediated transcription.29

To our knowledge, there are no published studies on EMSY in PrCa. Considering the location of EMSY and its connection to the DNA repair pathway, we screened EMSY for variants in Finnish familial and unselected PrCa patients and evaluated the association of the variants with PrCa risk.

Material and methods

Patients and controls

The clinical characteristics [mean age at diagnosis, Gleason grade and prostate specific antigen (PSA) value] of PrCa patients analyzed in this study are presented in Table 1. We screened the youngest affected patients from 184 Finnish PrCa families for EMSY sequence variants. All families had three or more affected members. The EMSY gene was also screened in 188 unselected Finnish PrCa patients (age at diagnosis < 62 years). After the initial screen, a total of seven sequence variants were validated in a larger sample set, which included 2,113 unselected PrCa cases (no age criteria) (Table 2). Aggressive unselected cases were defined based on a PSA value of ≥ 20 μg/L or a pathologic Gleason grade of ≥ 7 at diagnosis. The remaining cases were considered non-aggressive. All unselected cases genotyped in this study were collected from the Pirkanmaa Hospital District and diagnosed in Tampere University Hospital from 1969 to 2009 (clinical data was missing for 34 cases). As controls, a total of 923 samples from anonymous Finnish male blood donors were obtained from the Finnish Red Cross. All control individuals were healthy at the time of blood donation, and the blood samples were collected from the same geographical region as the unselected cancer cases. Full informed consent concerning the samples and patient information was obtained from all patients. The study was performed with the appropriate research permissions from the Ethics Committee of the Tampere University Hospital, Finland, and the Ministry of Social Affairs and Health in Finland.

Genotyping

Genomic DNA was extracted from peripheral blood leukocytes with the Puregene kit, according to the manufacturer’s instructions (Gentra Systems, Inc., Minneapolis, MN). All coding regions and exon-intron boundaries were amplified from the genomic DNA by polymerase chain reaction (PCR). All primer sequences are available upon request. The PCR products were purified with AcroPrep™ 96 (350 μl) Multi-
<table>
<thead>
<tr>
<th>Nucleotide change</th>
<th>Amino acid change</th>
<th>Unselected cases</th>
<th>Familial cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genotype distribution(^1), n (%)</td>
<td>MAF</td>
<td>Genotype distribution(^1), n (%)</td>
<td>MAF</td>
</tr>
<tr>
<td>IVS1-1890C&gt;T</td>
<td>IVS1-1890C&gt;T</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>IVS6-43A&gt;G</td>
<td>IVS6-43A&gt;G</td>
<td>2281-20-0 (99.1-0.9-0)</td>
<td>0.004</td>
<td>181-3-0 (98.4-1.6-0)</td>
</tr>
<tr>
<td>774T&gt;C</td>
<td>774T&gt;C</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>182-2-0 (98.9-1.1-0)</td>
</tr>
<tr>
<td>IVS8+40A&gt;G (rs4245443)</td>
<td>IVS8+40A&gt;G (rs4245443)</td>
<td>23-75-90 (12.2-39.9-47.9)</td>
<td>0.322</td>
<td>23-85-76 (12.5-46.2-41.3)</td>
</tr>
<tr>
<td>IVS8-60G&gt;C (rs74680029)</td>
<td>IVS8-60G&gt;C (rs74680029)</td>
<td>184-4-0 (97.9-2.1-0)</td>
<td>0.011</td>
<td>180-4-0 (97.8-2.2-0)</td>
</tr>
<tr>
<td>1215A&gt;G</td>
<td>1215A&gt;G</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>IVS9+17T&gt;A</td>
<td>IVS9+17T&gt;A</td>
<td>187-1-0 (99.5-0.5-0)</td>
<td>0.003</td>
<td>184-0-0 (100-0-0)</td>
</tr>
<tr>
<td>IVS10+37T&gt;A(C)</td>
<td>IVS10+37T&gt;A(C)</td>
<td>180-8-0 (95.7-4.3-0)</td>
<td>0.021</td>
<td>175-9-0 (95.1-4.9-0)</td>
</tr>
<tr>
<td>IVS10+68C&gt;G</td>
<td>IVS10+68C&gt;G</td>
<td>96-70-22 (51.1-37.2-11.7)</td>
<td>0.303</td>
<td>81-84-19 (44.0-45.7-10.3)</td>
</tr>
<tr>
<td>IVS11+14C&gt;T (rs11600501)</td>
<td>IVS11+14C&gt;T (rs11600501)</td>
<td>184-4-0 (97.9-2.1-0)</td>
<td>0.011</td>
<td>180-4-0 (97.8-2.2-0)</td>
</tr>
<tr>
<td>1764A&gt;G</td>
<td>1764A&gt;G</td>
<td>187-1-0 (99.5-0.5-0)</td>
<td>0.003</td>
<td>184-0-0 (100-0-0)</td>
</tr>
<tr>
<td>IVS13+47delA (rs11363199)</td>
<td>IVS13+47delA (rs11363199)</td>
<td>101-70-17 (53.7-37.2-9.1)</td>
<td>0.277</td>
<td>88-80-16 (47.8-43.5-8.7)</td>
</tr>
<tr>
<td>2681T&gt;G</td>
<td>2681T&gt;G</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3205G&gt;A</td>
<td>3205G&gt;A</td>
<td>1173-1-0 (99.9-0.1-0)</td>
<td>4x10^-4</td>
<td>184-0-0 (100-0-0)</td>
</tr>
<tr>
<td>3222C&gt;G (rs35962163)</td>
<td>3222C&gt;G (rs35962163)</td>
<td>1169-5-0 (99.9-0.1-0)</td>
<td>4x10^-4</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3648T&gt;C</td>
<td>3648T&gt;C</td>
<td>549-515-110 (46.7-43.9-9.4)</td>
<td>0.313</td>
<td>88-81-15 (47.8-44.0-8.2)</td>
</tr>
<tr>
<td>3716G&gt;A</td>
<td>3716G&gt;A</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3716G&gt;A</td>
<td>3716G&gt;A</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3716G&gt;A</td>
<td>3716G&gt;A</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3716G&gt;A</td>
<td>3716G&gt;A</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3716G&gt;A</td>
<td>3716G&gt;A</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
<tr>
<td>3716G&gt;A</td>
<td>3716G&gt;A</td>
<td>188-0-0 (100-0-0)</td>
<td>–</td>
<td>183-1-0 (99.5-0.5-0)</td>
</tr>
</tbody>
</table>

\(^1\)In the genotype distribution, three numbers separated by hyphens represent homozygotes, heterozygotes and homozygotes in the order which is based on nucleotide change marked on the name of the variant. \(^\ast\)MAF (minor allele frequency) ≥ 0.01 and not reported in the NCBI SNP database. \(^\ast\)denotes that the frequency of the variant was not determined.
well filter plates (PALL Life Sciences, Ann Arbor, MI) and sequenced using the BigDye Terminator Cycle Sequencing Reaction Kit v3.1 (Applied Biosystems, Foster City, CA), according to the manufacturer’s instructions. The sequencing reactions were purified by ethanol precipitation and sequenced with an ABI Prism Sequencer 3130xl (Applied Biosystems). PrCa samples with sequence variants were sequenced bidirectionally. Sequences were analyzed with the Sequencher™ (v4.7) (Gene Codes Corporation, Ann Arbor, MI). The observed variants were named based on a reference sequence, NM_020193. Sequences and detailed PCR and sequencing conditions are available upon request.

Three variants were validated by the available TaqMan® SNP Genotyping Assay, according to the manufacturer’s instructions using the ABI Prism® 7900HT (Applied Biosystems). Four other variants were validated using the high resolution melt (HRM) analysis on a Bio-Rad platform (Bio-Rad Laboratories, Hercules, CA). Detailed HRM reaction conditions, primer sequences and melting curves are available upon request. The variants were confirmed by sequencing.

Statistical methods
A χ²-test was performed to test each sequence variant locus for Hardy-Weinberg equilibrium (HWE) in controls and PrCa cases. The genotype frequencies were compared between controls and PrCa cases by calculating odds ratios (ORs) with 95% confidence intervals (CIs). ORs were calculated for each genotype separately because the dominant-recessive relationship between the alleles is currently not known. Association with aggressive PrCa was tested using the logistic regression, based on the OR values and 95% CIs (data not shown). The frequencies of the variants in familial PrCa, unselected PrCa and control samples are presented in Table 2.

The frequencies of 16 of the variants were determined in familial PrCa, unselected PrCa and control samples. After screening, we chose seven EMSY variants (IVS6-43A>G, 3205G>A, 3222C>G, 3250T>C, 3648T>C, IVS21+343A>G and IVS21+938A>G) for validation in the NCBI SNP database35 (rs entries) (Table 2). Most of the variants (n = 25) were SNPs; however, we also detected a rare, one-nucleotide insertion at the 3' UTR and a common intronic deletion. Of the variants with MAF ≥ 0.01, three were novel and had not been reported in the NCBI SNP database35 (Table 2). The frequencies of 16 of the variants were determined in control samples. After screening, we chose seven EMSY variants (IVS6-43A>G, 3205G>A, 3222C>G, 3250T>C, 3648T>C, IVS21+343A>G and IVS21+938A>G) for validation, based on the OR values and 95% CIs (data not shown). The frequencies of the variants in familial PrCa, unselected PrCa and control samples are presented in Table 2.

The distribution of the observed variants is shown in Figure 1. There were no variants in the ENT domain coding region or in the region encoding the HP1 and BS69 interacting region. There were no missense mutations in the exons that encode the BRCA2 interacting region of the protein, but two synonymous mutations and four intronic variants were found. Most of the observed variants were in the 3' region of the gene, specifically in the two last exons and the 3' UTR.

A novel intronic variant, IVS6-43A>G, was significantly associated with PrCa risk (Table 3 and Table 4). The association data (ORs and 95% CIs) for the other EMSY variants are not shown, because they were not statistically significant.

![Figure 1. Distribution of EMSY variants.](image-url)
All observed genotypes were in HWE in PrCa cases and controls \((p \geq 0.05)\). The IVS6-43A>G genotype carriers displayed an increased risk for familial PrCa \((OR = 7.5; 95\% CI = 1.3–45.5; p = 0.003)\) (Table 3). The \(p\) value exceeded the significance level when adjusted for multiple testing \((p = 0.02)\).

The IVS6-43A>G variant exhibited a marginally different frequency between unselected PrCa cases and controls \((OR = 3.5; 95\% CI = 0.8–15.2; p = 0.04)\), but it did not remain statistically significant after correction for multiple testing \((p = 0.36)\) (Table 3). When evaluating the significance of the association between IVS6-43A>G and cancer aggressiveness using case–control model, the variant was associated with aggressive unselected cases \((OR = 6.0; 95\% CI = 1.3–26.4; p = 0.008)\) (Table 4). Furthermore, when patients with aggressive or nonaggressive PrCa were directly compared (case–case analysis), the AG genotype was associated with aggressive PrCa \((OR = 6.5; 95\% CI = 1.5–28.4; p = 0.008)\) (Table 3). When evaluating the significance of the association between IVS6-43A>G and cancer aggressiveness using case–control model, the variant was associated with aggressive unselected cases \((OR = 6.0; 95\% CI = 1.3–26.4; p = 0.008)\) (Table 4). Furthermore, when patients with aggressive or nonaggressive PrCa were directly compared (case–case analysis), the AG genotype was associated with aggressive PrCa \((OR = 6.5; 95\% CI = 1.5–28.4; p = 0.008)\) (Table 3). After adjusting the \(p\) values for multiple testing, both remained statistically significant \((p = 0.03\) and \(p = 0.002\), respectively). No difference was found when nonaggressive cases were compared with controls \((OR = 0.5; 95\% CI = 0.04–5.9; p = 0.6)\).

In the study of familial PrCa, 3 of the 184 screened families had the PrCa risk-increasing genotype IVS6-43AG (observed in the youngest affected patient) (Table 2). Two of these families had samples available for further examination (Fig. 2). In family 359, three affected men (II-5, II-6, and III-1), who all had aggressive disease, had the PrCa risk-increasing allele G, and the nonaggressive case (II-1) was AA. Their sister (II-3), who has breast cancer, also had the associated variant, while a healthy sister (II-4) was wild type. Additionally, samples were available from two unaffected sons (III-2 and III-3) of II-6, and both had the PrCa associated genotype AG. In family 017, two men (II-5 and III-3), who both had aggressive PrCa, had the G variant. Two unaffected brothers (III-2 and III-5) of III-3 also had the variant, and one unaffected brother (III-1) was wild type. Both sons (IV-1 and IV-2) of III-3 and two out of four sons (III-10 and III-11) of II-5 carried the AG genotype. Samples were not available for two affected men (III-6 and III-7) in family 017, and both had the nonaggressive disease. All deceased aggressive PrCa affected men that had the IVS6-43AG genotype died of PrCa in these families.

To better understand the relationship between the\(\text{EMSY}\) variants, an LD analysis was conducted using Haploview (v4.1).\(^{32}\) Pairwise LD \(D^\prime\) and \(r^2\) values for the five validated variants and the haplotype block structures are presented in Supporting Information Figure 1. The LD analysis revealed one haplotype block that was composed of the SNPs 3648T>C, IVS21+343A>G and IVS21+938A>G. None of the haplotypes were associated with PrCa (data not shown). Based on the \textit{in silico} analysis, the IVS6-43A>G variant does not affect intrinsic regulatory elements, splicing elements, transcriptional regulatory motifs, or miRNA target sites.

### Table 3. Association of IVS6-43A>G with familial prostate cancer

<table>
<thead>
<tr>
<th>Genotype distribution of IVS6-43A&gt;G</th>
<th>OR (95% CI)</th>
<th>(p)</th>
<th>(p^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unselected cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2097-16-0</td>
<td>3.5 (0.8–15.2)</td>
<td>0.04</td>
<td>0.36</td>
</tr>
<tr>
<td>Familial cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>181-3-0</td>
<td>7.5 (1.3–45.5)</td>
<td>0.003</td>
<td>0.02</td>
</tr>
</tbody>
</table>

\(^1\)Adjusted for multiple testing.

| Screening cases \((n = 188)\) have been excluded from unselected cases. Abbreviations: OR = odds ratio; CI = confidence interval. |

### Table 4. Association of IVS6-43A>G with aggressive unselected prostate cancer

<table>
<thead>
<tr>
<th>Unselected cases</th>
<th>n</th>
<th>Genotype distribution of IVS6-43A&gt;G</th>
<th>MAF</th>
<th>OR (95% CI)</th>
<th>(p)</th>
<th>(p^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control design:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggressive</td>
<td>1013</td>
<td>1000-13-0</td>
<td>0.006</td>
<td>6.0 (1.3–26.4)</td>
<td>0.008</td>
<td>0.03</td>
</tr>
<tr>
<td>Nonaggressive</td>
<td>857</td>
<td>856-1-0</td>
<td>6 × 10(^{-4})</td>
<td>0.5 (0.04–5.9)</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Case-Case design:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggressive versus nonaggressive cases</td>
<td></td>
<td></td>
<td></td>
<td>6.5 (1.5–28.4)</td>
<td>0.003</td>
<td>0.002</td>
</tr>
</tbody>
</table>

\(^1\)Adjusted for multiple testing.

| Screening cases \((n = 188)\) have been excluded from unselected cases. Abbreviations: OR = odds ratio; CI = confidence interval; MAF = minor allele frequency. |
The most significant finding in this study was the association of a novel intronic variant, IVS6-43A>G, with PrCa. According to our results, the variant is associated with PrCa among familial patients and increases the risk for aggressive unselected PrCa. Our findings differ from the other studies that reported a lack of association of PrCa risk-increasing SNPs at 11q13 with aggressive PrCa. Interestingly, the SNP observed here is located at 11q13.5, about 7 Mb from the previously studied 11q13 SNPs. Despite their relatively close locations, these SNPs might be situated in different haplotype blocks; therefore, no LD would be observed between them. This lack of LD would explain the lack of association signal in the previous studies. On the other hand, our results are consistent with the pooled genome-wide linkage scan of the International consortium for prostate cancer genetics (ICPCG), which connected the locus on 11q14 to aggressive PrCa, particularly in pedigrees with an average age at diagnosis of 65 years or less.

All three variant-positive cases from the screened families had aggressive disease outcomes. However, the association with aggressive PrCa in families was not evaluated statistically because the small sample size did not have statistical power. We were able to study two of the IVS6-43A>G variant-positive families in more detail. Our finding that the variant segregated with aggressive PrCa in these families supports the aggressive PrCa promoting effect of IVS6-43A>G, which was first seen in the unselected cases. Interestingly, the variant also segregated with breast cancer, when the sisters with and without the disease in family 359 were considered. The unaffected variant-positive sons (III-2 and III-3) of II-6 in family 359 and sons (III-10, III-11, IV-1 and IV-2) of II-5 and III-3 in family 017 are still relatively young (the average age at diagnosis for PrCa in Finland is 72 years), which does not exclude future PrCa diagnoses. The present PrCa statuses of the two variant-positive brothers (III-2 and III-5) of III-3 in family 017 are unknown because the latest update available from the Finnish Cancer Registry was from 2009.

Current diagnostic methods, such as PSA screening, are not ideal for the discrimination of aggressive PrCa from the overwhelming majority of indolent cases. Therefore, aggressive PrCa associated biomarkers, such as genetic variants, are urgently needed, but they are thought to be rare in the population, which was observed for the IVS6-43A>G SNP in this study. Because no single SNP will likely be able to reliably distinguish between different types of PrCa, the prognosis is more likely to be based on a joint effect of multiple risk factors. The IVS6-43A>G variant, as presented here, was associated with aggressive PrCa, which suggests that this SNP is a good candidate for a future diagnostic SNP panel.

Despite multiple genome-wide studies (linkage and GWAS), no major risk genes for PrCa have been identified. This may be, at least in part, due to rare regulatory defects, which are difficult to identify and verify. The IVS6-43A>G variant is situated in the sixth intron of EMSY, 43 bp upstream of an exon, and its function is unclear. Based on its location, it might affect splicing of the EMSY mRNA, but no splicing elements were detected at the site using an in silico analysis. It is also possible that the observed association of the SNP is due to LD with a more distant functional variant, which needs to be further evaluated.

Figure 2. Segregation of IVS6-43A>G variant in two prostate cancer families. Squares denote male subjects, and circles signify female subjects. Black symbols indicate patients with prostate cancer, and symbols with black circle in the middle signify other cancers (Family 359: II-3 breast cancer, II-8 kidney cancer; Family 017: I-2 liver cancer, II-6 undefined cancer). IVS6-43AG carriers are denoted by a plus sign (+), and a minus sign (-) signifies IVS6-43AA genotype. Age at diagnosis for affected and current age for unaffected persons in years (Y) is presented below the symbol. The age at death in years (d. Y) is shown below symbols for deceased family members. Aggressive cases were defined based on a prostate specific antigen (PSA) value of ≥ 20 µg/L or a pathologic Gleason grade of ≥ 7 at diagnosis.
In conclusion, we report the novel variant IVS6-43A>G that is associated with aggressive unselected PrCa and contributes to familial cancer risk in a Finnish population. These findings indicate that the genetic mechanisms for aggressive and nonaggressive PrCa are different and that the IVS6-43A>G SNP is a potential variant that could be used as a genetic biomarker for aggressive PrCa in predictive medicine. This finding warrants functional studies and studies in other populations.

Acknowledgements

The authors thank Riitta Vaalavuo, Linda Enroth and Aleksandra Bebel for assistance. The prostate cancer patients and their families are thanked for their participation and cooperation. This study was financially supported by the Competitive Research Funding of the Tampere University Hospital (grant 91.091), the Reino Lahtikari Foundation, the Finnish Cancer Organizations, the Sigrid Juselius Foundation and the Academy of Finland (126/41 for T.W. and 116437 for J.S.).
identifies a subset of prostate cancer with an aggressive behavior. *Prostate*, in press.


30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007;81:559–75.


Fine mapping of 11q13.5 identifies regions associated with prostate cancer and prostate cancer death

Riikka Nurminen, Rainer Lehtonen, Anssi Auvinen, Teuvo L.J. Tammela, Tiina Wahlfors, Johanna Schleutker

Institute of Biomedical Technology/BioMediTech and Prostate Cancer Research Center, University of Tampere and Finlab Laboratories, Biokatu 8, FI-33014 Tampere, Finland
Department of Biosciences, University of Helsinki, PO BOX 65, FI-00014 Helsinki, Finland
School of Health Sciences, University of Tampere, Medisinuksenkatu 3, FI-33014 Tampere, Finland
Department of Urology and Prostate Cancer Research Center, University of Tampere and Tampere University Hospital, Teiskontie 35, PO BOX 2000, FI-33521 Tampere, Finland
Department of Medical Biochemistry and Genetics, University of Turku, Kiiinamyllynkatu 10, FI-20014 Turku, Finland

Available online 2 July 2013

KEYWORDS
Genetic association study
EMSY
Death
Prostate cancer
Single nucleotide polymorphism

Abstract Background: Chromosomal region 11q13–14 associates with prostate cancer (PrCa). Previously, we identified a rare intronic mutation on EMSY (11q13.5) that increases the risk of aggressive PrCa and associates with familial PrCa. Here, we further study the genetic structure and variants of the PrCa susceptibility region 11q13.5.

Methods: This study included 2716 unselected hospital-based PrCa cases, 1318 cases of a screening trial and 908 controls of Finnish origin. We imputed single nucleotide polymorphisms (SNPs) and structural variants from the 1000 Genomes Project and validated the associations of the variants in two PrCa patient sets by genotyping. Genetic structure was studied with haplotype analysis.

Results: Two independent regions at 11q13.5 were associated with PrCa risk. The most significant association was at EMSY (rs10899221, odds ratio (OR) 1.29–1.40, P = 3.5 × 10^{-4}–0.002) near the previously identified mutation. Correlated intronic SNPs rs10899221 and rs72944758 formed with other EMSY variants common and rare haplotypes that were associated with increased risk (P = 4.0 × 10^{-4}) and decreased risk (P = 0.01) of PrCa, respectively. The other associated region was intergenic. Among the six validated variants, rs12277366 was...
1. Introduction

Hereditary factors increase the risk of prostate cancer (PrCa), with a heritability estimate of 16–45%.[1] Genome-wide association studies (GWAS) and fine-mapping studies have identified multiple single nucleotide polymorphisms (SNPs) that associate with PrCa predisposition. However, GWAS SNPs only account for a portion of the estimated heritability.[2,3]

Chromosomal region 11q is linked with hereditary PrCa in the Finnish population.[4] The most significant linkage signal was detected between 11q13.4 and 11q22 (D11S1314 and D11S898), with the highest peak at 11q14.1 (D11S901). An association of 11q13 with PrCa was further supported by GWAS studies that identified two correlated SNPs at 11q13.2 (rs7931342 and rs10896449).[5,6] We previously screened genetic variants in exons and exon–intron boundaries of EMSY (C11orf30, 11q13.4–11q13.5) in Finnish PrCa patients.[7] EMSY interacts with BRCA2 and chromatin remodelling proteins[8] and regulates the transcription of interferon-stimulated genes.[9] Our study identified a rare intronic mutation IVS6-43A>G (rs200331695) that increases the risk of aggressive PrCa and associates with familial PrCa in the Finnish population.[7] No causative genetic variants known to contribute to cancer development have been identified in the 11q13–14 locus.

In this study, we fine mapped the chromosomal region 11q13.5 around the previously identified mutation to characterise the genomic structure of 11q13.5 and to identify the PrCa predisposing genetic variants in this region.

2. Patients and methods

2.1. Study population

The clinical characteristics of PrCa patients are summarised in Table 1. A total of 2716 unselected hospital-based patients of Finnish origin were identified from the Pirkanmaa Hospital District. Furthermore, the study included 1318 PrCa patients diagnosed in the Finnish component of the European Randomized Study of Screening for Prostate Cancer described in detail elsewhere.[10] The patients were from the screening arm of the prostate-specific antigen (PSA) screening trial. A total of 908 control samples originated from healthy Finnish male blood donors provided by the Finnish Red Cross. Full written informed consent concerning the samples and patient information was obtained from the patients. The study was performed with the appropriate research permissions from the Ethics Committee of the Tampere University Hospital, Finland, and the Ministry of Social Affairs and Health in Finland.

2.2. Genotyping and statistical methods

Genotyping is described in Supplementary Methods. A total of 31 tagging SNPs (tagSNPs) were determined covering region 76154846–76569208 at 11q13.5 (Supplementary Methods and Supplementary Table 1) and were genotyped from 1811 unselected hospital-based PrCa patients and 908 controls. Prior imputation samples with genotype call rates <90% (11/2719) and tagSNPs, which deviated from Hardy–Weinberg equilibrium (HWE) in the controls (1/31), were excluded from the study. Imputation was performed using IMPUTE2 v2.2.3[11,12] with default settings and a 1000 Genomes cosmopolitan reference panel.[13] Imputed variants with expected minor allele frequency (MAF) ≥1% and an info score ≥0.8 and variants with MAF <1% and an info score ≥0.9 were included in the association analysis. Five genetic models of association (additive, dominant, recessive, general and heterozygote) were tested using a Frequentist test, which accounts for imputation uncertainty, with the score method implemented in SNPTEST v2.3.0.[14] Variants deviating from HWE in the controls were excluded from the analysis (1/605).

Validated variants were tested for HWE in the controls and for an association with a risk of PrCa using Cochran–Armitage trend test. We used a case–case design to test the association of the validated variants with clinicopathological features of unselected PrCa [diagnosis age, Gleason score (aggressive cancer), PrCa death, PSA level and progression after prostatectomy, radiation therapy and hormone therapy]. We calculated the P values adjusted for correlated tests ($$P_{\text{ACT}}$$) to account for multiple testing.[15] Statistical analyses were performed using PLINK v1.07[16] and R v2.11.1.[17] Pairwise measures of linkage disequilibria (LD) between SNPs ($$r^2$$), haplotype-block structure based on Gabriel’s definition[18] and haplotype associations were calculated.
using Haploview v4.2. The results were considered significant when \( P < 0.05 \).

The effects of rs72944758 and rs10899221 on splicing were predicted with Human Splicing Finder v2.4.1. The functionality of all validated variants was predicted with HaploReg v2, which includes annotations from the Roadmap Epigenomics and the Encyclopedia of DNA Elements (ENCODE) data.

3. Results

3.1. Imputation

We imputed genotypes across a 900 kb region on 11q13.5. A total of 604 imputed genetic variants were tested for association. We observed association signals from two regions. The \(-\log_{10} P\) values of the additive model are presented for imputed variants and tagSNPs in Fig. 1. An imputed SNP, rs12277366, on an intergenic region showed the strongest evidence of association (\( P = 0.002 \)). From that region, we selected five SNPs and a one-nucleotide insertion for validation based on the following criteria: SNPs that were associated with PrCa in the additive model (\( P < 0.01 \)) or were in high LD with the aforementioned SNPs (\( r^2 \geq 0.8 \)) and were associated in any of the five genetic models (\( P < 0.01 \)). Another significant association signal was at tagSNP rs10899221 located in the intron of EMSY (\( P = 0.007 \)). The only flanking SNP in high LD (\( r^2 = 0.92 \)) with rs10899221 was an imputed EMSY SNP rs72944758. Both intronic SNPs were included in the validation.

3.2. Validation of association

A total of eight genetic variants were genotyped in two PrCa patient sets and controls. All variants were in HWE in the controls (\( P > 0.05 \)). The most significant association with a risk of PrCa was observed for the tag-SNP rs10899221 located in an intron of EMSY (\( P = 0.007 \)). The only flanking SNP in high LD (\( r^2 = 0.92 \)) with rs10899221 was an imputed EMSY SNP rs72944758. Both intronic SNPs were included in the validation.

Table 1

<table>
<thead>
<tr>
<th>Clinical characteristics of the prostate cancer patients.</th>
<th>Unselected patients</th>
<th>Patients of the PSA screening trialb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (n)</td>
<td>2716a</td>
<td>1318</td>
</tr>
<tr>
<td>Diagnosis year (data missing, n)</td>
<td>1969–2009 (32)</td>
<td>1996–2009 (0)</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>69 (36–94; 32)</td>
<td>65 (54–78; 0)</td>
</tr>
<tr>
<td>PSA (µg/L) at diagnosis (n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;4)</td>
<td>184</td>
<td>165</td>
</tr>
<tr>
<td>4–9.9</td>
<td>976</td>
<td>837</td>
</tr>
<tr>
<td>10–19.9</td>
<td>666</td>
<td>204</td>
</tr>
<tr>
<td>(\geq 20)</td>
<td>659</td>
<td>97</td>
</tr>
<tr>
<td>Data missing</td>
<td>231</td>
<td>15</td>
</tr>
<tr>
<td>Biopsy Gleason grade at diagnosis (n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\leq 6)</td>
<td>1002</td>
<td>372</td>
</tr>
<tr>
<td>7</td>
<td>608</td>
<td>131</td>
</tr>
<tr>
<td>(\geq 8)</td>
<td>380</td>
<td>62</td>
</tr>
<tr>
<td>Data missing or other than biopsy</td>
<td>726</td>
<td>753</td>
</tr>
<tr>
<td>Prostate cancer death (n)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Died of cancer</td>
<td>423</td>
<td>40</td>
</tr>
<tr>
<td>Non-deceased, follow-up (&gt;6) years</td>
<td>1141</td>
<td>888</td>
</tr>
<tr>
<td>PSA progression (n out of n treated)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostatectomy</td>
<td>256/736 (35%)</td>
<td>79/259 (31%)</td>
</tr>
<tr>
<td>Hormone therapy</td>
<td>378/875 (43%)</td>
<td>20/68 (29%)</td>
</tr>
<tr>
<td>Radiation therapy</td>
<td>159/368 (43%)</td>
<td>63/157 (40%)</td>
</tr>
</tbody>
</table>

Abbreviation: PSA, prostate-specific antigen.  
\( a \) Includes 1811 unselected prostate cancer patients of imputation.  
\( b \) Prostate cancer patients of the screening arm of the PSA screening trial described in Schröder et al.  

R. Nurminen et al. / European Journal of Cancer 49 (2013) 3335–3343 3337
that was more common in cases (26.0%) than controls (21.4%, $P = 4.0 \times 10^{-4}$) (Fig. 2). In addition, a rare haplotype was identified with the risk alleles in block 1, but the haplotype was more common in the controls (2.3%) than in the cases (1.4%, $P = 0.01$) (Fig. 2).

The five closely positioned SNPs (rs12271056, rs12277366, rs2155225, rs2155222, rs7126722) and a one-nucleotide insertion (rs143975731) were associated with a risk of PrCa in the unselected cases (Table 2). The only SNP that was significant in both patient sets after the correction for multiple testing was rs12277366 (unselected cases OR 1.15, $P_{ACT} = 0.03$; PSA screening trial cases OR 1.17, $P_{ACT} = 0.03$) (Table 2), which had the strongest signal also in the imputed data set. The SNP rs12277366 was in high LD with rs143975731, rs2155225 and rs2155222 ($r^2 = 0.85–0.99$) in the genotyped data set.

The haplotype analysis identified a PrCa-associated common haplotype in block 2 ($P = 0.02$) that included the risk alleles of rs12271056, rs143975731 and rs12277366 (Fig. 2). The haplotype frequencies for the cases and controls were 46.7% and 43.4%, respectively. SNPs rs2155225, rs2155222 and rs7126722 were located in a different haplotype block than the aforementioned intergenic SNPs. These SNPs formed a haplotype (block 4) with non-risk alleles that was more common in the controls (16.9%) than in the cases (14.4%, $P = 0.02$) (Fig. 2).

In addition to the overall risk of PrCa, all validated intergenic variants were associated with PrCa death (OR 0.72–0.80, $P_{ACT} = 2.4 \times 10^{-4}$–0.02) (Table 3). The alleles predisposed to PrCa were observed more frequently in non-deceased patients (follow-up >6 years) than in patients with a lethal disease (Table 3). In addition, three of the variants (rs143975731, rs12277366, rs2155225) associated with Gleason score with a similar pattern, but the associations did not remain significant after correction for multiple testing (Table 3).

**EMSY** SNPs, rs10899221 and rs72944758, which showed the strongest association with PrCa after validation, were predicted to be located on potential splice sites that were affected by the nucleotide change (Table 4). In addition, both SNPs were predicted to brake or introduce a motif of a splicing silencer protein hnRNPA1 and to affect the exonic splicing enhancer (ESE) elements of SR proteins (Table 4).

The functional annotations of the validated variants are presented in Supplementary Table 3. The **EMSY** SNPs rs10899221 and rs72944758 are located within open chromatin, and rs10899221 is within a binding site of transcription factor CEBPB. Both SNPs were predicted to affect regulatory motifs with varying impacts. The locations of the three intergenic validated variants (rs12271056, rs143975731, rs12277366) coincided with the chromatin states of enhancers based on both the Roadmap Epigenomics and the ENCODE data. A significant enrichment of cell type-specific (umbilical vein endothelial cells and B-lymphocytes) enhancers was observed (Supplementary Table 4). In addition,
multiple regulatory motifs were predicted to be affected by the intergenic variants with varying impacts (Supplementary Table 3).

4. Discussion

We conducted a fine-mapping study to investigate the contribution of the chromosomal region 11q13.5 to PrCa predisposition. Two correlated EMSY SNPs, rs72944758 and rs10899221, were associated with the disease and formed a haplotype that was also associated with an increased risk of PrCa. In addition, a set of correlated intergenic SNPs showed similar associations. Yet, the EMSY SNPs and the intergenic SNPs were in different haplotype blocks separated by recombination hot spots (Figs. 1 and 2), which may indicate that these two regions contribute independently to disease risk.

The PrCa risk associated haplotype of EMSY included the major alleles of rs72944758 and rs10899221, which increased the risk of PrCa based on the single SNP association analyses. Although the frequencies of the risk alleles of rs72944758 and rs10899221 were high in the population (92.0% and 89.0%, respectively), the phenotypic effect size may be lower because of the lower frequency of the susceptibility haplotype (26.0%). Interestingly, the alleles were included in a different, rare haplotype (frequency of 1.4%) that was associated with a decreased risk of PrCa. This indicates that genetic interactions between SNPs within a gene may contribute to the phenotype; i.e. an effect of a SNP is regulated by another SNP in the same haplotype. In fact, genetic interactions have been suggested to explain at least a proportion of the missing heritability of common diseases.

The PrCa risk associated haplotype of EMSY included the major alleles of rs72944758 and rs10899221 located in consecutive introns of EMSY. Interestingly, the aggressive PrCa associated mutation rs200331695, which we had identified previously, is located in the same intron as rs72944758. No moderate or high LD was observed between rs200331695 and the imputed or genotyped SNPs, which is expected because typically rare variants

### Table 2
Association of validated SNPs and a one-nucleotide insertion with the risk of prostate cancer.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Positiona</th>
<th>Allelesb</th>
<th>Study set</th>
<th>Genotype distributionc</th>
<th>OR (95% CI)</th>
<th>( P^d )</th>
<th>( P_{\text{ACT}}^e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs72944758</td>
<td>76173948</td>
<td>A/G</td>
<td>Unselected cases</td>
<td>2295/383/15</td>
<td>1.26 (1.04–1.52)</td>
<td>\textbf{0.02}</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>1139/162/6</td>
<td>1.46 (1.17–1.83)</td>
<td>7.2 \times 10^{-4}</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>734/151/9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10899221</td>
<td>76181631</td>
<td>G/A</td>
<td>Unselected cases</td>
<td>2151/507/36</td>
<td>1.29 (1.10–1.52)</td>
<td>\textbf{0.002}</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>1062/231/15</td>
<td>1.40 (1.16–1.69)</td>
<td>3.5 \times 10^{-4}</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>670/218/12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12271056</td>
<td>76349466</td>
<td>T/C</td>
<td>Unselected cases</td>
<td>1447/1036/182</td>
<td>1.19 (1.06–1.34)</td>
<td>\textbf{0.004}</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>676/537/87</td>
<td>1.13 (0.99–1.29)</td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>432/393/70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs143975731</td>
<td>76349715-</td>
<td>( \sim /G )</td>
<td>Unselected cases</td>
<td>593/1322/765</td>
<td>1.14 (1.03–1.27)</td>
<td>\textbf{0.01}</td>
<td>0.03</td>
</tr>
<tr>
<td>rs143975716</td>
<td>76349716</td>
<td>G/A</td>
<td>PSA screening cases</td>
<td>283/653/362</td>
<td>1.15 (1.02–1.30)</td>
<td>\textbf{0.02}</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>171/436/288</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12277366</td>
<td>76350119</td>
<td>G/C</td>
<td>Unselected cases</td>
<td>596/1328/768</td>
<td>1.15 (1.03–1.28)</td>
<td>\textbf{0.01}</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>289/657/359</td>
<td>1.17 (1.04–1.32)</td>
<td>\textbf{0.01}</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>174/434/293</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2155225</td>
<td>76351897</td>
<td>A/T</td>
<td>Unselected cases</td>
<td>596/1323/757</td>
<td>1.13 (1.01–1.26)</td>
<td>\textbf{0.03}</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>289/658/354</td>
<td>1.15 (1.02–1.30)</td>
<td>\textbf{0.02}</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>174/440/281</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2155222</td>
<td>76353805</td>
<td>G/A</td>
<td>Unselected cases</td>
<td>713/1314/648</td>
<td>1.15 (1.03–1.28)</td>
<td>\textbf{0.01}</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>326/663/311</td>
<td>1.12 (0.99–1.26)</td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>205/446/245</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7126722</td>
<td>76354161</td>
<td>G/A</td>
<td>Unselected cases</td>
<td>1457/1046/179</td>
<td>1.17 (1.04–1.32)</td>
<td>\textbf{0.008}</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSA screening cases</td>
<td>677/540/85</td>
<td>1.11 (0.97–1.27)</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Controls</td>
<td>439/395/67</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bold signifies \( P < 0.05 \).

**Abbreviations:** SNP, single nucleotide polymorphism; OR, odds ratio; CI, confidence interval.

a Genome Build 37.

b The first nucleotide is the risk allele.

c In the genotype distribution, three numbers separated by hyphens represent homozygotes, heterozygotes and homozygotes in the order which is based on the alleles.

d Cochran–Armitage trend test.

e \( P \) value adjusted for multiple testing using \( P \) values adjusted for correlated tests (\( P_{\text{ACT}} \)).15
do not exhibit LD with either rare or common SNPs due to low allele frequency. In silico analyses indicated that rs72944758 and rs10899221 may affect splicing motifs.

Table 3
Association of validated genetic variants \((n = 8)\) with aggressive prostate cancer defined by Gleason (Gl) score \((\text{cases} = \text{Gl} \geq 8; \text{controls} = \text{Gl} \leq 6)\) and with prostate cancer death \((\text{cases} = \text{died of prostate cancer}; \text{controls} = \text{non-deceased with follow-up} \geq 6\text{years})\). Patients with intermediate Gleason score 7 were excluded due to a high chance of misclassification between aggressive and non-aggressive diseases. The associations were tested for prostate cancer risk increasing alleles of the variants using unselected patients.

<table>
<thead>
<tr>
<th>Clinicopathological feature</th>
<th>Genetic variant</th>
<th>Frequency cases</th>
<th>Frequency controls</th>
<th>OR (95% CI)</th>
<th>(P^a)</th>
<th>(P_{\text{ACT}}^b)</th>
</tr>
</thead>
</table>
| Gleason score               | rs72944758_A   | 0.918          | 0.922             | 0.94 (0.69–1.28) | 0.7  | \n | rs10899221_G                | 0.894          | 0.888             | 1.06 (0.81–1.39) | 0.7  | \n | rs12271056_T                | 0.727          | 0.738             | 0.95 (0.78–1.14) | 0.6  | \n | rs143975731_-               | 0.433          | 0.481             | 0.82 (0.70–0.98) | 0.03 | 0.09 \n | rs12277366_G                | 0.433          | 0.483             | 0.82 (0.69–0.97) | 0.02 | 0.08 \n | rs2155225_A                 | 0.432          | 0.483             | 0.81 (0.68–0.96) | 0.02 | 0.06 \n | rs2155222_G                 | 0.485          | 0.525             | 0.85 (0.72–1.01) | 0.07 | \n | rs7126722_G                 | 0.729          | 0.739             | 0.95 (0.78–1.15) | 0.6  | \n | Prostate cancer death       | rs72944758_A   | 0.919          | 0.928             | 0.88 (0.65–1.18) | 0.4  | \n | rs10899221_G                | 0.891          | 0.899             | 0.92 (0.71–1.18) | 0.5  | \n | rs12271056_T                | 0.697          | 0.751             | 0.76 (0.64–0.91) | \n | rs143975731_-               | 0.403          | 0.481             | 0.73 (0.62–0.86) | 0.003 | \n | rs12277366_G                | 0.400          | 0.482             | 0.72 (0.61–0.84) | 4.8 \times 10^{-4} | \n | rs2155225_A                 | 0.404          | 0.484             | 0.72 (0.62–0.85) | 8.2 \times 10^{-4} | \n | rs2155222_G                 | 0.467          | 0.524             | 0.80 (0.68–0.94) | 0.006 | 0.02 \n | rs7126722_G                 | 0.692          | 0.752             | 0.74 (0.62–0.88) | 7.8 \times 10^{-4} | 0.003 \n
|                    |                |                |                  |              |      | \n
Bold signifies \(P < 0.05\).

Abbreviations: OR, odds ratio; CI, confidence interval.

\(^a\) Cochran–Armitage trend test.

\(^b\) \(P\) value adjusted for multiple testing using \(P\) values adjusted for correlated tests \((P_{\text{ACT}})\).
In addition, the exon codes the region surrounding the aberrant pre-mRNA processing of EMSY may result in defects in protein–protein interactions and the disturbance of downstream functions.

The intergenic SNPs (including a one-nucleotide insertion) were associated with PrCa in the unselected cases, but for five out of six variants, the association was not replicated in the PrCa patients of the PSA screening trial. The cancer status of the trial participants was screened at 4 years intervals, which has been reported to result in diagnosed cancers that would not have caused clinical symptoms during men’s lifetimes. The overdiagnosed cases in the sample set could explain the observed lack of association of the intergenic SNPs with a risk of PrCa.

One of the most interesting findings was the association of the intergenic SNPs with PrCa death. Unexpectedly, the PrCa predisposing alleles decreased the risk, and the non-predisposing alleles increased the risk of PrCa deaths. A similar pattern was observed for aggressive cancer. A comparable finding has been reported for PrCa deaths. A similar pattern was observed for aggressive cancer.

The closest transcribed gene from the associated region is LRRC32 (~15 kb). LRRC32 is a transmembrane protein that is specific to regulatory T cells and essential for TGF-β expression. In addition, large intergenic non-coding RNAs (lincRNAs) are located in the interferon response. An aberrant pre-mRNA processing of EMSY may result in defects in protein–protein interactions and the disturbance of downstream functions.

---

**Table 4**  
*In silico* prediction of the effects of rs72944758 and rs10899221 on splicing motifs.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Nucleotide change</th>
<th>Potential splice site&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Branch point</th>
<th>Enhancer motif&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Silencer motif&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs72944758</td>
<td>G&gt;A</td>
<td>Site broken (72.26 &gt; 61.68)</td>
<td>New site</td>
<td>0 &gt; 80.78</td>
<td>(0 &gt; 73.57)</td>
</tr>
<tr>
<td>rs10899221</td>
<td>A&gt;G</td>
<td>New site (60.61 &gt; 71.19)</td>
<td>Site broken</td>
<td>Three sites broken (77.63 &gt; 0; 72.31 &gt; 0; 80.96 &gt; 0)</td>
<td>(67.62 &gt; 0)</td>
</tr>
</tbody>
</table>

<sup>a</sup> The consensus values (CV) of motifs are reported in brackets. Only motifs with the relative change of CV >10% are reported. Strong site CV ≥ 80 and less strong site 70 ≤ CV < 80.

<sup>b</sup> Exonic splicing enhancer (ESE) motifs of SR proteins [SRp40, SC35, SF2/ASF, SF2/ASF (IgM/BRCA1), SRp55]. Motifs with relative change of CV = 100% are reported.

<sup>c</sup> Silencer motifs of hnRNP proteins.

---

Abbreviation: SNP, single nucleotide polymorphism.

---

disease-causing mutation on MTRR, which activated an ESE and led to aberrant splicing because an intronic sequence was included within the mRNA. The exon between introns containing rs72944758 and rs10899221 codes for amino acid Ser209, which is phosphorylated in the interferon response. The phosphorylation by Akt1 relieves the EMSY-mediated transcriptional repression of target genes. In addition, the exon codes for a part of the EMSY protein that intersects with BRCA2. BRCA2 confers increased susceptibility not only to PrCa development but also to more aggressive disease and poor prognosis. An aberrant pre-mRNA processing of EMSY may result in defects in protein–protein interactions and the disturbance of downstream functions.

The overdiagnosed cases in the sample set could explain the observed lack of association of the intergenic SNPs with PrCa. The overdiagnosed cases in the sample set could explain the observed lack of association of the intergenic SNPs with PrCa. The overdiagnosed cases in the sample set could explain the observed lack of association of the intergenic SNPs with PrCa.
remains unclear. This result warrants further study to determine the functionality and causality of the identified loci.

Conflict of interest statement

None declared.

Acknowledgements

We thank the prostate cancer patients for their participation in this study. We also thank Ekaterina Slitkova and Riina Liikala for their contribution to the genotyping and Riitta Vaalavuo for her assistance. The Epigenomics Consortium (http://commonfund.nih.gov/epigenomics) is acknowledged for the Roadmap data, and the ENCODE consortium is acknowledged for the ENCODE data. The study was supported by the Competitive Research Funding of the Tampere University Hospital (Grants 9L091 and 9M094), the Finnish Cancer Organisations, the Sigrid Juselius Foundation and the Academy of Finland (Grants 116437 and 251074). The study funding sources had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejca.2013.06.006.

References


