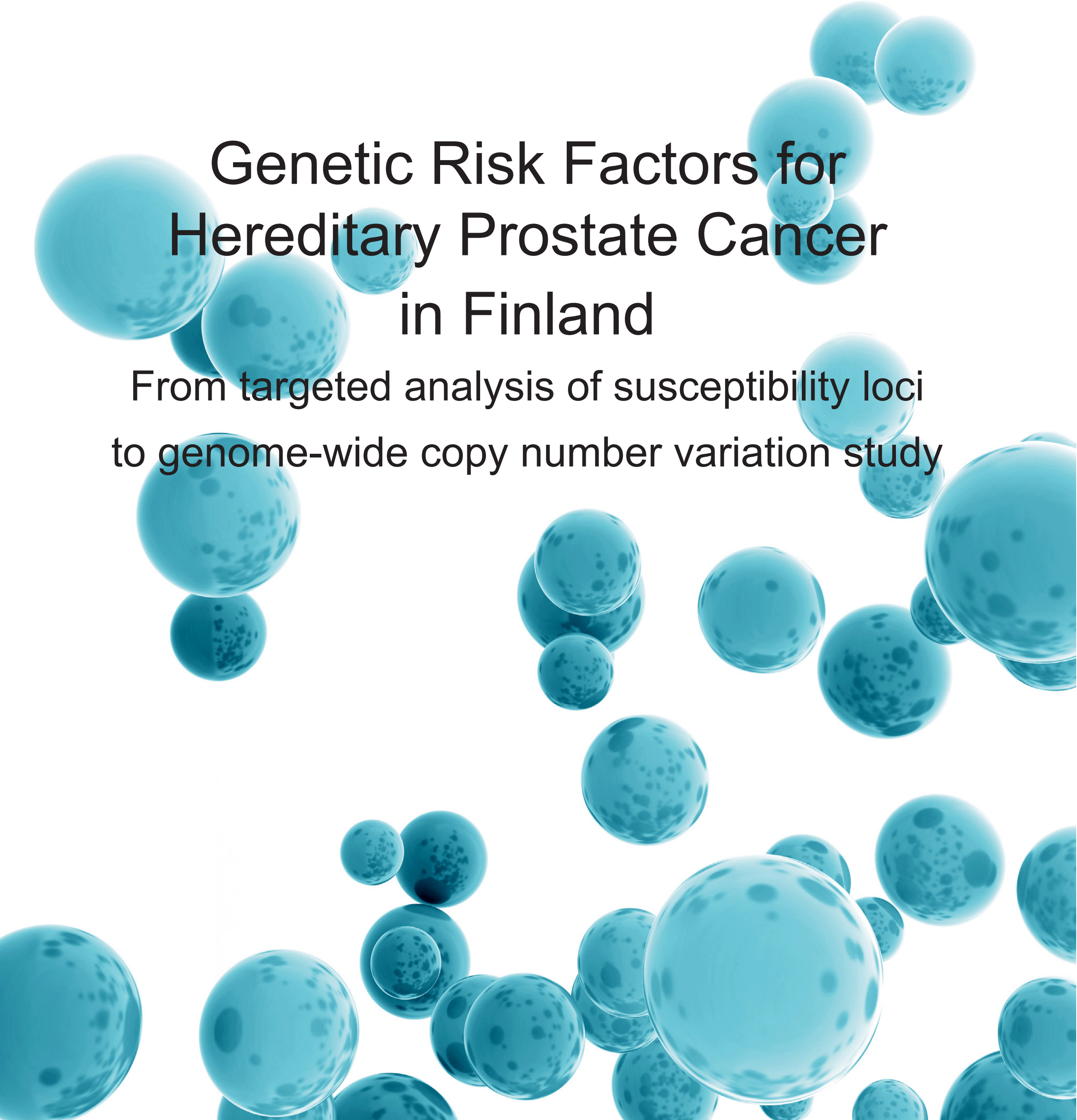
Acta Universitatis Tamperensis 2179

VIRPI LAITINEN

Genetic Risk Factors for Hereditary Prostate Cancer in Finland

From targeted analysis of susceptibility loci to genome-wide copy number variation study



VIRPI LAITINEN

Genetic Risk Factors for Hereditary Prostate Cancer in Finland

From targeted analysis of susceptibility loci to genome-wide copy number variation study

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 17 June 2016, at 12 o'clock.

UNIVERSITY OF TAMPERE

VIRPI LAITINEN

Genetic Risk Factors for Hereditary Prostate Cancer in Finland

From targeted analysis of susceptibility loci to genome-wide copy number variation study

Acta Universitatis Tamperensis 2179
Tampere University Press
Tampere 2016

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere, BioMediTech
Laboratory of Cancer Genetics
Fimlab Laboratories
Finland

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

Reviewed by
Docent Markku Vaarala
University of Oulu
Finland
Docent Elisabeth Widén
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.

Cover design by
Mikko Reinikka

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

Acta Universitatis Tamperensis 2179
ISBN 978-952-03-0143-9 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1678
ISBN 978-952-03-0144-6 (pdf)
ISSN 1456-954X
http://tampub.uta.fi

Suomen Yliopistopaino Oy – Juvenes Print
Tampere 2016

YMPÄRISTÖMERKKI
MILJÖMÄRKT
441 729
Painotuote

Contents

List of Original Communications

This thesis is based on the following communications, referred in the text by their Roman numerals (I-III). In addition, some unpublished results are presented.

I **Laitinen VH***, Wahlfors T*, Saaristo L, Rantapero T, Pelttari LM, Kilpivaara O, Laasanen S-L, Kallioniemi A, Nevanlinna H, Aaltonen L, Vessella RL, Auvinen A, Visakorpi T, Tammela TLJ, Schleutker J (2013). *HOXB13* G84E mutation in Finland: Population-based analysis of prostate, breast and colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev* 22(3):452-460. *equal contribution

II **Laitinen VH**, Rantapero T, Fischer D, Vuorinen EM, Tammela TLJ, PRACTICAL Consortium, Wahlfors T, Schleutker J (2015). Fine-mapping the 2q37 and 17q11.2-q22 loci for novel genes and sequence variants associated with a genetic predisposition to prostate cancer. *Int J Cancer* 136(10):2316-2327.

III **Laitinen VH**, Akinrinade O, Rantapero T, Tammela TLJ, Wahlfors T, Schleutker J (2016). Germline copy number variation analysis in Finnish families with hereditary prostate cancer. *Prostate* 76(3):316-324.

The original publications have been reproduced with the permission of the copyright holders.

Abbreviations

BPH	Benign Prostatic Hyperplasia
cDNA	Complementary DNA
CI	Confidence Interval
CNV	Copy Number Variant/Variation
COSMIC	Catalogue of Somatic Mutations in Cancer
CRPC	Castration-Resistant Prostate Cancer
DDPC	Dragon Database of Genes Implicated in Prostate Cancer
DE	Differentially Expressed (Genes)
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic Acid
dsDNA	Double-Stranded DNA
EMBL-EBI	European Bioinformatics Institute (part of the European Molecular Biology Laboratory)
ENCODE	The Encyclopedia of DNA Elements
eQTL	Expression Quantitative Trait Locus/Loci
ERSPC	The European Randomized Study of Screening for Prostate Cancer
ExAC	Exome Aggregation Consortium
FFPE	Formalin-Fixed and Paraffin-Embedded (Tissue)
FIMM	Institute for Molecular Medicine Finland
GO	Gene Ontology
GWAS	Genome-Wide Association Study
HGMD	Human Gene Mutation Database
HLOD	Heterogeneity Logarithm of Odds
HPC	Hereditary Prostate Cancer
HR	Hazard Ratio
HWE	Hardy-Weinberg Equilibrium
iCOGS	International Collaborative Oncological Gene-Environment Study

KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
lincRNA	Long Noncoding RNA
LOD	Logarithm of Odds
LOH	Loss of Heterozygosity
MAF	Minor Allele Frequency
MALDI-TOF	Matrix-Assisted Laser Desorption Ionization Time-of-Flight (Technology)
miRNA	MicroRNA
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NHGRI	National Human Genome Research Institute
OMIM	Online Mendelian Inheritance in Man
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PIA	Proliferative Inflammatory Atrophy
PIN	Prostate Intraepithelial Neoplasia
PON-P	Pathogenic-Or-Not Pipeline
PSA	Prostate Specific Antigen
qPCR	Quantitative (Real-Time) PCR
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
SBS	Sequencing-By-Synthesis
SNP	Single Nucleotide Polymorphism
SUMO	Small Ubiquitin-like Modifier
TF	Transcription Factor
TSS	Transcription Start Site
VCP	Variant Calling Pipeline
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

Abstract

Prostate cancer is the most frequently diagnosed male malignancy in industrialized Western countries. In Finland, approximately 5000 new cases emerge each year, which is equal to more than one-third of all male cancers. Following lung cancer, prostate cancer is the second most common cause of cancer death in Finland (Finnish Cancer Registry). The burden not only to the patients and their families but also to the national health care system is, therefore, significant.

While the etiology of prostate cancer is not yet fully understood, a few specific risk factors have been recognized, including advanced age, ethnic origin and positive family history. In addition to genetic predisposition, environmental factors, diet and hormones likely modify the disease risk. A majority of prostate cancer cases are sporadic, but approximately 5-10% of cases can be classified as hereditary cancers, which result from inherited germline variants predisposing their carriers to the disease. In prostate cancer, genetic factors play an essential role and have been estimated to explain as much as 58% of the cancer risk. Unlike other common cancers, such as breast or colorectal cancer, prostate cancer is genetically very heterogeneous, which has made the identification of genetic susceptibility factors extremely challenging. Only a few high-risk candidate genes and variants have been found, and the risk effect of the more common variants is typically low. As a consequence, in many Finnish prostate cancer families, the underlying causative gene defects remain unknown.

The aim of this thesis study was to identify novel genetic factors contributing to prostate cancer predisposition in Finland. The search focused especially on two chromosomal regions, 2q37 and 17q11-q22, which have repeatedly shown a strong linkage with increased prostate cancer risk in various populations. These two loci were characterized by sequencing samples representing both familial and unselected prostate cancer patients, as well as unaffected controls. In addition, a genome-wide copy number variation analysis was performed on familial prostate cancer patients to locate genomic alterations associated with increased risk of hereditary prostate cancer.

Additional evidence for a role in prostate carcinogenesis was obtained for several previously reported candidate genes, including *HOXB13* and *ZNF652* at 17q21.3

and *HDAC4* and *ANO7* at 2q37. In particular, the importance of the *HOXB13* variant p.G84E was established in this study. This variant was observed at a frequency of 8.4% among familial prostate cancer patients (vs. 1.0% in controls), making it the most common prostate-cancer-associated risk variant detected in Finland thus far. This variant was also associated with earlier age at disease onset (<55 years). In a sequence analysis, potential risk alleles were identified in the other candidate genes as well. Two *ZNF652* variants and one *HDAC4* variant were shown to associate significantly with hereditary prostate cancer. Although the co-occurrence of these variants with the disease was incomplete, the variants were more common among prostate cancer patients than among unaffected family members. The sequencing of the coding region of the *ANO7* gene revealed eight possibly pathogenic variants, but additional co-segregation and association analyses are required to establish their clinical significance.

In addition, novel putative prostate cancer candidate genes were identified, most importantly *EPHA3* at 3p11.1. The *EPHA3* gene codes for a receptor tyrosine kinase that is responsible for signal transduction between neighbouring cells. This gene is commonly mutated in several cancers. In this study, a 14.7 kb intronic deletion within the *EPHA3* gene was detected in 11.6% of familial prostate cancer patients but in only 6.1% of unaffected controls. The results also suggest that *EPHA3* deletion may predispose patients to a more aggressive form of the disease, but this finding requires further validation.

In this thesis, several hereditary factors likely contributing to prostate cancer susceptibility were identified in previously reported and novel prostate cancer candidate genes. These findings need to be confirmed in further studies. It is possible, however, that in the future, some of the observed variants may be applied in clinical diagnostics, for example, for the early identification of individuals with high prostate cancer risk.

Tiivistelmä

Eturauhassyöpä on miesten yleisin syöpäsairaus teollistuneissa länsimaissa, myös Suomessa. Maassamme tehdään vuosittain noin 5000 uutta eturauhassyöpädiagnoosia ja tautia sairastaa tälläkin hetkellä noin 47000 miestä. Eturauhassyövän osuus kaikista miehillä todetuista syövistä on noin kolmannes, ja keuhkosyövän jälkeen se on toiseksi yleisin miesten syöpäkuolemien aiheuttaja maassamme (Suomen Syöpärekisteri). Sairauden kansanterveydellinen taakka on siis merkittävä.

Eturauhassyövän etiologiaa ei kuitenkaan vielä tarkkaan tunneta. Kyseessä on monitekijäinen sairaus, jonka keskeisimpiin riskitekijöihin kuuluvat yli 55 vuoden ikä, etninen tausta sekä positiivinen sukuanamneesi. Myös ympäristötekijät, ruokavalio ja hormonit saattavat vaikuttaa sairastumisriskiin. Vaikka valtaosa eturauhassyövistä on sporadisia eli satunnaisia, voidaan syöpä noin 5-10 %:ssa tapauksista luokitella perinnölliseksi. Perinnöllisessä syöpäalttiudessa potilas on perinyt toiselta tai molemmilta vanhemmiltaan yhden tai useamman geenivirheen, jotka lisäävät syöpään sairastumisen riskiä merkittävästi. Eturauhassyövässä perinnöllisillä tekijöillä on poikkeuksellisen vahva rooli ja niiden on arvioitu selittävän jopa 58 % eturauhassyöpäriskistä. Useista kattavista tutkimuksista huolimatta eturauhassyövän taustalta on kyetty tunnistamaan vain muutamia korkean riskin alttiusgeenejä. Näiden lisäksi on löydetty useita, suhteellisen yleisiä matalan riskin variantteja, jotka lisäävät syöpäriskiä vain hieman. Monista muista yleisistä syövistä, kuten rinta- tai kolorektaalisyövästä poiketen eturauhassyöpä onkin geneettisesti hyvin heterogeeninen, minkä seurauksena riskiyksilöiden tunnistaminen ja taudin vaikeusasteen varhainen ennustaminen on hyvin haastavaa.

Aiemmissa tutkimuksissa on toistuvasti havaittu kromosomialueiden 2q37 ja 17q11-q22 yhteys kohonneeseen eturauhassyöpäriskiin. Väitöskirjatyössä näiltä kytkentäalueilta etsittiin sekvensoimalla geenivirheitä, jotka liittyvät erityisesti perinnölliseen eturauhassyöpään suomalaisväestössä. Lisäksi eturauhassyöpäsukujen potilailta kartoitettiin kopioluvun muutoksia koko genomin alueelta, ja selvitettiin, assosioituvatko ne eturauhassyöpään suomalaisessa perheaineistossa.

Kytkentäalueilla sijaitsee useita eturauhassyöpäalttiuteen liitettyjä geenejä, kuten *HOXB13* ja *ZNF652* lokuksessa 17q21.3 sekä *HDAC4* ja *ANO7* lokuksessa 2q37. Väitöskirjatyössä näiden geenien rooli eturauhassyövän kandidaattigeeneinä

vahvistui edelleen. Keskeisimmäksi osoittautui *HOXB13*-geeni ja erityisesti sen p.G84E-variantti. Tutkimuksessa havaittiin, että perinnöllistä eturauhassyöpää sairastavista potilaista peräti 8.4 % oli variantin kantajia, verrokeista vain 1.0 %. p.G84E-variantti on siis toistaiseksi yleisin suomalaispotilailla todettu eturauhassyövälle altistava geenivirhe. Lisäksi todettiin, että variantin kantajilla oli kohonnut riski sairastua eturauhassyöpään alle 55-vuotiaana. Sekvenssianalyysissä myös muista kandidaattigeeneistä tunnistettiin muutoksia, jotka saattavat altistaa kantajansa eturauhassyövälle. *ZNF652*-geenissä todettiin kaksi varianttia ja *HDAC4*-geenissä yksi variantti, jotka assosioituivat merkitsevästi perinnölliseen eturauhassyöpään. Vaikka variantit eivät segregoituneet perheissä täydellisesti yhdessä taudin kanssa, olivat ne selkeästi yleisempiä syöpäpotilailla kuin terveillä perheenjäsenillä. *ANO7*-geenin sekvenssianalyysissä tunnistettiin kahdeksan mahdollisesti patogeenista varianttia, mutta näiden varianttien kliinisen merkityksen selvittäminen edellyttää jatkotutkimuksia.

Väitöskirjatyössä löydettiin myös uusia mahdollisia eturauhassyövän kandidaattigeenejä, joista tärkeimpänä kromosomialueella 3p11.1 sijaitseva *EPHA3*. *EPHA3* koodaa reseptorityrosiinikinaasia, joka osallistuu solujen väliseen signaalinvälitykseen. Geenin mutaatioita on todettu useissa eri syöpätyypeissä. Tässä tutkimuksessa havaittiin, että *EPHA3*-geenin introniin paikantuva, noin 14.7 kiloemäksen (kb) deleetio oli lähes kaksi kertaa yleisempi eturauhassyöpäpotilailla (kantajafrekvenssi 11.6 %) kuin verrokeilla (6.1 %). Lisäksi saatiin viitteitä siitä, että *EPHA3*-deleetion kantajilla saattaa olla kohonnut riski sairastua taudin aggressiiviseen muotoon. Tämän tuloksen vahvistaminen vaatii vielä lisätutkimuksia.

Väitöskirjatyössä jo aiemmin raportoiduista alttiusgeeneistä sekä uusista kandidaattigeeneistä tunnistettiin siis useita perinnöllisiä geenivirheitä, jotka saattavat altistaa kantajansa eturauhassyövälle. Löydösten kliininen merkitys tulee vielä varmentaa jatkotutkimuksissa. On kuitenkin mahdollista, että tulevaisuudessa joitakin nyt havaituista varianteista voidaan käyttää kliinisessä diagnostiikassa, esimerkiksi korkean syöpäriskin potilaiden varhaiseen tunnistamiseen.

1 Introduction

Cancer is a common disease that can develop in almost any human tissue. The estimated lifetime risk of cancer is approximately one in four. The disease is strongly associated with advanced age, with more than 90% of cancers being diagnosed among older adults (aged >45 years). In 2014, the most prevalent cancer types among Finnish men were prostate cancer, lung cancer and colon cancer. Two of the deadliest cancers, lung and prostate cancer, explained 35% of the cancer-specific mortality. Among women, breast cancer predominated, and it was the most commonly diagnosed cancer and the primary cause of cancer-related death (Finnish Cancer Registry).

Cancers are disorders that are characterized by uncontrolled cell proliferation. When normal cells gradually evolve towards malignancy, they acquire biological properties that enable tumour growth and metastasis. Typically, cancer cells are able to stimulate cell division, escape from growth suppressors, resist cell death (apoptosis), maintain replicative immortality, induce blood vessel formation (angiogenesis), and activate invasion and metastasis. Additional representative features include the ability to reprogramme energy metabolism and to avoid immune destruction (Hanahan & Weinberg 2011). A fully transformed cancer cell is immortal, resistant to most drugs and capable of spreading to nearby and distant tissues (Horne et al. 2015).

Several environmental and lifestyle factors, such as smoking, diet, infections, and exposure to ultraviolet light, ionizing radiation or pollution, have been listed as possible causes of cancer. However, fundamentally, cancer is a disease of the genome and results from genomic instability. Tumourigenesis is triggered by mutations in one or a few key genes known as gatekeepers or caretakers, which normally stabilize the genome. These mutations then allow the cell to outgrow its surrounding cells (Vogelstein et al. 2013). As cancer progresses, additional genomic rearrangements occur, leading to the accumulation of chromosomal deletions and translocations, as well as somatic mutations, which activate oncogenes and inactivate tumour suppressor genes. Together, these events explain the genetic heterogeneity observed in many human cancers (Horne et al. 2015).

In sporadic cancer, all mutations within a cell are somatic and will not be transmitted to the next generation. However, approximately 5-10% of cancer cases represent hereditary cancer, in which a mutation predisposing to the disease has been inherited from one of the parents. Carriers of such germline mutations are at an increased risk of developing cancer. The most common familial cancer types include breast, ovarian, colon and prostate cancers. Hereditary cancer may be suspected in a family with several affected first- or second-degree relatives, patients diagnosed at an early age or patients having multiple primary tumours (Cole et al. 1996). Similar molecular mechanisms are probably responsible for the development of hereditary and sporadic forms of cancer (Cussenot et al. 1998). Therefore, candidate genes identified in studies of hereditary cancer likely explain a proportion of sporadic cancers as well.

This study focused on elucidating the genetic changes predisposing to hereditary prostate cancer. Inherited factors are known to contribute significantly to this disease, and the most prominent individual risk factor is positive family history (Zeegers et al. 2003). However, the identification of risk genes and variants is a laborious process. During decades of intensive research, it has become evident that susceptibility to prostate cancer is more complex than initially presumed. Several different candidate genes have been found, illustrating the genetic heterogeneity and polygenic inheritance of the disease. The individual variants that confer high cancer risk are generally rare, whereas common variants increase the risk only slightly (Eeles et al. 2014). In addition, some disease-associated alleles show reduced penetrance, and the roles of copy number changes and regulatory variants are just beginning to emerge. Clinically, the severity of prostate cancer varies from indolent to aggressive, and in early stages of the disease, it may be difficult to recognize the patients at risk of lethal disease (Demichelis & Stanford 2015). The need for novel biomarkers enabling accurate diagnostics and personalized treatment strategies is therefore apparent. Improved prognosis is invaluable to cancer patients and their close relatives. Medical doctors treating the patients will benefit from clinical practice guidelines tailored according to the patient's genomic mutation profile. Furthermore, a deep knowledge of the genetic background of prostate cancer will be the key to the prevention of this common disease in the future.

2 Review of the Literature

All gene and protein names and symbols that appear in this thesis follow the nomenclature guidelines of the HUGO Gene Nomenclature Committee (HGNC; Wain et al. 2002).

2.1 Prostate cancer

In developed Western countries, including European countries, United States, Australia and New Zealand, the most common malignancy in men is prostate cancer. More than one million new diagnoses and >300,000 prostate-cancer-related deaths are reported worldwide each year (GLOBOCAN 2012).

In Finland, the incidence and prevalence of this disease are high and are expected to increase in the future due to the ageing of the population. Prostate cancer represents approximately one-third of all male cancers and is the second most common cause of cancer death. In 2014, a total of 4,596 new cases were diagnosed, 47,000 men were living with the disease and 856 men died of it. Most prostate cancers are non-aggressive, and the relative 5-year survival rate is as high as 93% (Finnish Cancer Registry).

2.1.1 Etiology and risk factors

Prostate cancer is a multifactorial disease that develops as a result of interplay between genetic, environmental and dietary factors (Bostwick et al. 2004). The most well-established risk factors include advanced age, ethnic background and a positive family history (Crawford 2003). In addition, the role of hormones and inflammation has been investigated, but their contribution to disease susceptibility is less clear.

2.1.1.1 Age and ethnicity

Prostate cancer affects predominantly men older than 40 years (Tao et al. 2015). Currently, the average age at diagnosis in Finland is 70 years, and only 4.4% of newly diagnosed patients are younger than 55 years (Finnish Cancer Registry). The lifetime risk to Finnish men of developing prostate cancer is 12.0% (Hjelmborg et al. 2014).

In addition to advanced age, ethnic origin influences prostate cancer risk. Even 25-fold differences in prostate cancer incidence have been reported worldwide (GLOBOCAN 2012). The disease is most common among Australian, New Zealand and African-American men, followed by Western and Northern Europeans (Center et al. 2012). In these countries, the high incidence is partially due to the high detection rate resulting from routine screening and diagnostics. Prostate cancer is also relatively common in the Caribbean, Southern Africa and South America. In contrast, in Eastern and South-Central Asia, the incidence of this disease is substantially lower (Center et al. 2012, GLOBOCAN 2012). Genetic factors likely explain a proportion of the observed variation. The severity of prostate cancer among black men born in the United States, Jamaica, West Africa and sub-Saharan Africa was evaluated in a recent study, and the results showed that the country of origin did not affect the clinical characteristics of the disease (Fedewa & Jemal 2013). Another study investigated the lifetime risk of prostate cancer among the major ethnic groups living in the United Kingdom, and striking differences between the groups were observed. Prostate cancer risk for black men was 1 in 4, for white men 1 in 8, and for Asian men 1 in 13 (Lloyd et al. 2015).

2.1.1.2 Family history

Many common cancers tend to cluster in families, and prostate cancer is no exception. Approximately 5-10% of prostate cancer cases represent familial cancers which are believed to result from heritable high-risk genetic factors (Carter et al. 1993). Several familial and epidemiological surveys have shown that in prostate cancer susceptibility, the effect of the genetic component is exceptionally strong (e.g., Steinberg et al. 1990, Carter et al. 1992, Grönberg et al. 1996, Hemminki & Czene 2002, Zeegers et al. 2003). In a large prospective study of Nordic twins, the cumulative incidence of prostate cancer was compared between monozygotic and dizygotic twin pairs. The results indicated that as much as 58% of prostate cancer risk is explained by genetic factors (Hjelmborg et al. 2014).

Prostate cancer risk correlates with the number of affected relatives. Sons and brothers of prostate cancer patients have a 2- to 4-fold increased cancer risk compared to that of the general population (Hemminki & Czene 2002, Zeegers et al. 2003, Kicinski et al. 2011). The age-specific hazard ratios (HRs), calculated using data stored in the Swedish population-based Family-Cancer Database, further illustrate the effect of family history on prostate cancer risk (Figure 1). For a man younger than 75 years, the HR of prostate cancer is 2.1 if only his father is affected, 3.0 if he has one affected brother and 8.5 if both his father and two brothers are affected. The highest HR of 17.7 is observed for men with three affected brothers (Brandt et al. 2010).

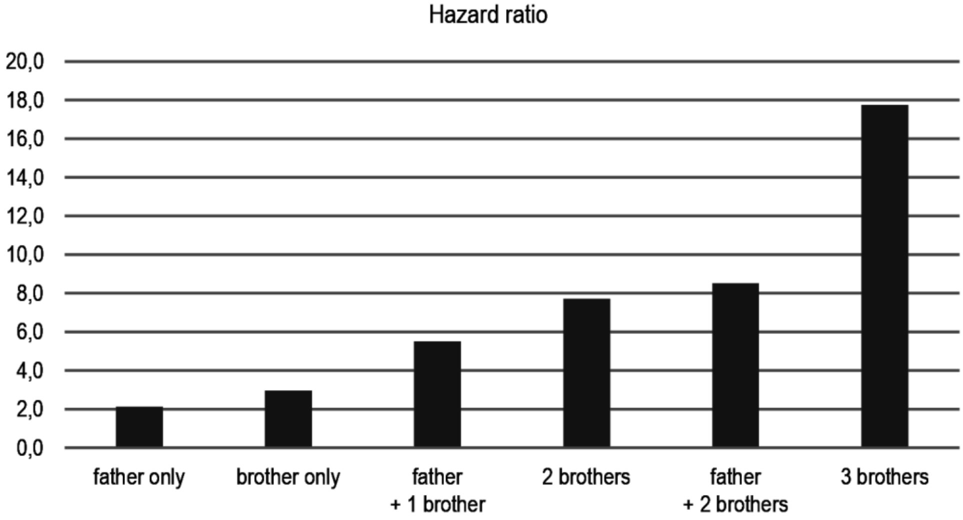


Figure 1. Hazard ratios for familial prostate cancer according to the number of affected relatives (modified from Hemminki 2012). The bar chart is based on the data published by Brandt et al. 2010.

The definition of hereditary prostate cancer (HPC) was introduced by Carter and colleagues in 1993 to aid in the collection of familial high-risk datasets that could then be used to map prostate cancer candidate genes. HPC refers to families that meet at least one of the following criteria: three or more first-degree relatives are affected with prostate cancer, prostate cancer is observed in three successive generations, or two first-degree relatives have been diagnosed with prostate cancer before the age of 55 years (Carter et al. 1993).

2.1.1.3 Environmental and dietary factors

The effect of diet on prostate cancer risk has been extensively studied, but the definitive link between dietary components and early stages of cancer remains unclear. Obesity is associated with increased risk of aggressive prostate cancer, prostate cancer recurrence and mortality (Allott et al. 2013). Negative effects have also been suggested for high-fat diets and for the consumption of well-cooked red meat (Hori et al. 2011), but the association is uncertain (Lin et al. 2015). In contrast, beneficial dietary factors include fruits and vegetables, especially tomatoes, which are rich in lycopene, as well as diets low in saturated fats and carbohydrates (Lin et al. 2015). Protective effects have also been reported for broccoli, soy, green tea and vitamin D (Schwartz 2014, Hackshaw-McGeagh et al. 2015). In addition, physical activity has been shown to slightly decrease prostate cancer risk (Liu et al. 2011).

The contribution of certain prostatic diseases to increased prostate cancer risk has been extensively investigated. Chronic inflammation certainly plays a role (Sfanos & De Marzo 2012), although the infectious micro-organism has not yet been identified. Possibly, the asymptomatic inflammatory process persists several years before cancer begins to develop (Sfanos et al. 2013). A few studies have reported an increased risk of prostate cancer for patients who have previously been diagnosed with benign prostatic hyperplasia (BPH) (Orsted et al. 2011, Saaristo et al. *unpublished results*). In addition, hormones, especially androgens, may be involved in prostate carcinogenesis by promoting the progression of the disease from the preclinical stage to the clinical stage (Bostwick et al. 2004). According to a recently proposed model, low levels of testosterone disturb androgens and androgen receptor (AR) signalling (Zhou et al. 2015). In addition, dietary oestrogens have been suggested to damage the prostate epithelium, thus leading to inflammation and increased cancer risk (Nelson et al. 2014).

2.1.2 Clinical characteristics

The prostate is an oval-shaped exocrine gland that belongs to the male reproductive and urinary tracts. It is located in front of the rectum and below the urinary bladder. An average adult prostate is approximately the size of a walnut and weighs 15-20 grams, but the size varies from man to man and tends to increase with age. The major function of the prostate is to produce seminal fluid, but it also participates in controlling urine flow (Bhavsar & Verma 2014). More than 95% of prostate cancers are adenocarcinomas originating from the prostatic epithelium (Shen & Abate-Shen 2010). Adenocarcinoma refers to a cancer that begins in the secretory cells of an internal gland. Typically, prostate carcinomas are multifocal (Villers et al. 1992). Primary tumours have been shown to contain several independent cancer foci that represent different genotypes (Bostwick et al. 1998, Macintosh et al. 1998). Metastases can have either monoclonal or polyclonal origins. Monoclonal metastases arise from a single ancestral cell present in the primary tumour (Liu et al. 2009a), whereas polyclonal metastases originate from several distinct subclones and, hence, reflect greater genomic diversity (Gundem et al. 2015).

The first precursor lesion observed in prostatic epithelial cells is a PIN (prostatic intraepithelial neoplasia), a condition where the structure and function of the epithelial cells has become abnormal. A low-grade PIN is usually harmless, whereas most patients with a high-grade PIN develop prostate cancer within the next ten years (Bostwick & Cheng 2012). A finding similar to a PIN is proliferative inflammatory atrophy (PIA), which can be observed in the prostate epithelium due to inflammation. This lesion is generally regarded as benign (Woenckhaus & Fenic 2008).

The clinical course of prostate cancer is highly variable, ranging from indolent, slow-growing and localized tumours to aggressive, fast-growing tumours that may metastasize to bones, lymph nodes or visceral organs, such as the liver. Usually, prostate cancer develops slowly with a long, asymptomatic preclinical phase. The first clinical symptoms are similar to those observed in BPH, including inflammation of the prostate gland, urethritis, bladder dysfunction, obstruction of the urethra and/or increased frequency of urination, especially at night. Advanced prostate cancer can cause haematuria, impotence and pains in different areas of the body, often due to bone metastases. With the exception of an earlier age of onset, the clinical features of hereditary prostate cancer do not differ from those of sporadic prostate cancer (Schaid 2004).

2.1.3 Diagnostics and screening

If prostate cancer is suspected, the initial scan includes an evaluation of prostate size and consistency by digital rectal examination and/or measurement of the prostate-specific antigen (PSA) concentration in the serum. PSA is a glandular serine protease that is produced and secreted by the epithelial cells of the prostate. It is encoded by the *Kallikrein-Related Peptidase 3 (KLK3)* gene. In prostate cancer, the normal epithelium is damaged, and an increased amount of PSA is released into blood circulation (Stamey et al. 1987). The cut-off values for normal total PSA levels depend on age and range from <2.5 ng/ml for men in their 40s to <6.5 ng/ml for men in their 70s (Oesterling et al. 1993). However, an elevated PSA value can also indicate benign conditions, such as BPH or prostatitis. Therefore, the total serum PSA level gives only an estimate of the likelihood of cancer. Generally, PSA values between 4 and 10 ng/ml predict the risk of prostate cancer to be approximately 25%, but if the total PSA is higher than 10 ng/ml, the risk of cancer is greater than 50% (Greene et al. 2013).

When abnormal results are obtained in the initial scan, a prostate biopsy is needed to confirm (or exclude) the diagnosis of cancer. Tumour tissue observed in the histopathological analysis of the biopsy sample is graded using the Gleason scoring system, which evaluates the level of cancer cell differentiation and aggressiveness (Epstein et al. 2016). Two of the most predominant tissue patterns are graded from 1 to 5 and are summed to calculate the Gleason score. Gleason scores ≥7 indicate a biologically aggressive cancer (Greene et al. 2013). Primary tumours are also classified according to the TNM (tumour, node, metastasis) staging system (Cheng et al. 2012), where T denotes the size and the invasiveness of the tumour (T1-4), N reveals whether the disease has spread to the regional lymph nodes (N0 or N1), and M describes distant metastasis (M0 or M1). TNM staging aids in treatment planning and in the estimation of prognosis.

Recently, the use of PSA-based screening in the detection of prostate cancer has become a controversial issue. The reported advantages include reduced prostate cancer specific mortality (Schröder et al. 2009) and earlier diagnosis (Kilpeläinen et al. 2010). However, as a consequence, the numbers of unnecessary biopsies and of the overdiagnosis and overtreatment of indolent cancers have increased, especially in older men (Schröder et al. 2009). Reported estimates of overdiagnosis range from 27% to 60% for cancers detected by screening (Sandhu & Andriole 2012). To improve the benefit-to-harm ratio, it has been suggested that screening should be focused on younger men (≤60 years) and that screening of older men should be restricted to only those with PSA values clearly above the threshold at the initial screening (Loeb et al. 2012, Vickers et al. 2014).

To complement the currently used screening and detection strategies, several clinical testing laboratories have introduced genetic tests aimed at identifying mutations in prostate-cancer-associated genes. According to the Orphanet (www.orpha.net/) and GeneTests (www.genetests.org/) websites, more than 20 molecular genetic tests for prostate cancer are now commercially available in several European and Northern American countries. Approximately half of these tests are multigene panels, containing 13 to 94 genes, and are designed to assess the genetic predisposition for up to nine hereditary cancers. The tests specific for familial prostate cancer are listed in Table 1.

Table 1. Commercially available genetic tests (n = 16) for familial prostate cancer.

Test	Method	Gene(s)*	Laboratory
Prostate cancer sequencing panel	Sequencing	*BRCA1, BRCA2, CHEK2, NBN, TP53*	CEN4GEN (Edmonton, Canada)
Molecular diagnosis of familial prostate cancer	Full gene sequencing, Deletion/Duplication testing	*BRCA2, ELAC2, RNASEL, SRD5A2, STAG1, ZNF783*	Centogene AG (Rostock, Germany)
Prostate cancer test	Sequencing	*BRCA2, ELAC2, RNASEL, SRD5A2*	Diagenom GmbH (Rostock, Germany)
Molecular diagnosis of familial prostate cancer	NA	*BRCA2*	Institut für Klinische Genetik (Stuttgart, Germany)
PCA3 for prostate cancer	Mutation scanning of select exons	*PCA3*	Parseh Pathobiology & Genetics Laboratory (Tehran, Iran)
Molecular diagnosis of familial prostate cancer	NA	*HOXB13*	Azienda Ospedaliera Istituti Ospitalieri di Cremona (Cremona, Italy)
HOXB13 gene analysis	Sequencing	*HOXB13*	Academic Medical Centre (Amsterdam, Netherlands)
Prostate cancer test (genetic predisposition)	Sequencing, Deletion/ Duplication testing	*CHEK2, NBN*	GENESIS Center for Medical Genetics (Poznan, Poland)
Prostate cancer 1	Sequencing	*RNASEL*	CGC Genetics (Porto, Portugal)
Molecular diagnosis of familial prostate cancer	Sequencing	*BRCA2, CHEK2*	CIALAB (Alicante, Spain)
Molecular diagnosis of familial prostate cancer	Sequencing	*BRCA1, BRCA2*	Lorgen G.P. (Armilla, Spain)
Molecular diagnosis of familial prostate cancer	Sequencing	*BRCA1, BRCA2, CHEK2*	Laboratorio de Genética Clinica S.L. (Madrid, Spain)

Table 1. Continued.

Test	Method	Gene(s)*	Laboratory
Molecular diagnosis of susceptibility to familial prostate cancer	Sequencing, MLPA	*BRCA1, BRCA2*	IMOMA (Oviedo, Spain)
Molecular diagnosis of predisposition to breast and prostate cancer	NA	*BRCA1, BRCA2*	Genetiks – Genetic diagnosis and research centre (Istanbul, Turkey)
Molecular diagnosis of *HNF1B*-gene-related diseases	Sequencing, MLPA	*HNF1B*	Centre Hospitalier Universitaire Vaudois (Lausanne, Switzerland)
HOXB13 mutation analysis (G84E)	Sequencing	*HOXB13*	Mayo Clinic (Rochester, USA)

** BRCA1/2 = Breast Cancer 1/2 Early Onset, CHEK2 = Checkpoint Kinase 2, ELAC2 = ElaC Ribonuclease Z 2, HNF1B = Hepatocyte Nuclear Factor 1-Beta, HOXB13 = Homeobox B13, NBN = Nibrin, PCA3 = Prostate Cancer Associated 3, RNASEL = Ribonuclease L, SRD5A2 = Steroid-5-Alpha-Reductase 2, STAG1 = Stromal Antigen 1, TP53 = Tumour Protein 53, ZNF783 = Zinc Finger Family Member 783.* NA = not available.

2.1.4 Medical therapies

Several different treatment options for prostate cancer exist, and the choice of strategy depends on the severity of the symptoms as well as the clinical and pathological characteristics of the tumour. Active surveillance is sometimes sufficient for localized, indolent cancers, especially if the patient is older than 70 years and has additional diseases or if the tumour is small in size and grows slowly. More aggressive cancers that have not spread into nearby tissues or lymph nodes and have not metastasized (T1-2, N0, and M0) are generally treated by radical prostatectomy or radiation therapy, which can be either external or internal (brachytherapy). These can be complemented with hormonal androgen-deprivation therapy (Attard et al. 2016). Less frequent approaches include, for example, cryotherapy and High-Intensity Focused Ultrasound (Autran-Gomez et al. 2012).

Unfortunately, curative treatment for advanced, metastatic prostate cancer (T1-4, N0-1, and M1) is not yet available. Disease progression can be delayed by surgical or chemical castration and by using anti-androgens, in combination with radiation and chemotherapy (Attard et al. 2016). Despite treatment, metastatic disease usually develops into castration-resistant prostate cancer (CRPC). The median overall survival time of men diagnosed with metastatic prostate cancer is approximately 42 months. CRPC diagnosis shortens the median overall survival time dramatically to only 18 months (James et al. 2015).

2.2 Cancer genetics

Cancer is a genetic disorder. The transformation of a cell from benign to malign arises from genomic instability, which leads to the accumulation of mutations in the genome of the cell. Typically, this process includes multiple steps and lasts for decades (Isaacs & Kainu 2001). Mutations that alter the expression of genes responsible for cell division, growth, differentiation or apoptosis provide the cell with a selective growth advantage that usually results in tumour formation (Vogelstein et al. 2013). Eventually, the tumour invades surrounding tissues and metastasizes to distant organs. Most tumours are monoclonal, originating from a single mother cell.

The key mutations steering tumourigenesis are called driver mutations. In common solid tumours, two to eight driver mutations are required to trigger the neoplastic process (Vogelstein et al. 2013). Additional, usually dozens but occasionally even hundreds of thousands of mutations may be present in the same cell, but these passenger mutations do not contribute to disease pathogenesis. Characteristic, frequently observed alterations of cancer cell genomes include mutations in oncogenes, tumour suppressor genes and DNA repair genes (Isaacs & Kainu 2001). The classification of cancer-related genes into these three subgroups is not always straightforward, as some genes display both oncogenic and tumour-suppressing features, while others exert their tumour suppressor properties via DNA repair. In addition, epigenetic alterations modify the expression of these genes, adding another level of complexity to the function of cancer genomes.

2.2.1 Oncogenes

Proto-oncogenes control normal cell proliferation. Typically, proteins encoded by proto-oncogenes function as growth factors, growth factor receptors, tyrosine kinases, signal transduction molecules, transcription factors (TFs) or anti-apoptotic molecules. An activating gain-of-function mutation may transform the proto-oncogene into an oncogene that can induce malignant growth. Mutations that activate oncogenes are dominant at the cellular level and include point mutations (usually missense mutations), amplifications and chromosomal rearrangements, resulting in gene fusions or up-regulated oncogene expression (Todd & Wong 1999).

Several oncogenes involved in prostate carcinogenesis have been identified. The translocation of the 5' untranslated region of *TMPRSS2 (Transmembrane Protease Serine 2)* to *ERG (V-Ets Avian Erythroblastosis Virus E26 Oncogene Homolog)*, a TF belonging to the *ETS* family of oncogenes, is found in approximately 50% of prostate cancer samples (Tomlins et al. 2005). The *TMPRSS2-ERG* fusion results in the androgen-regulated overexpression of truncated ERG protein (Clark et al. 2007) and has been reported to be associated with poor prognosis in localized cancer (Demichelis et al. 2007). The amplification of the *MYC* locus at 8q24 is observed in 2-20% of prostate cancers (Khemlina et al. 2015). The *MYC* gene (*V-Myc Avian Myelocytomatosis Viral Oncogene Homolog*) codes for a TF involved in cell cycle progression, apoptosis and cellular transformation (Grandori et al. 2000). Another frequent alteration is the overexpression of androgen receptor (AR), which can result from gene amplification, point mutations or altered splicing (Visakorpi et al. 1995). AR is a steroid-hormone-activated TF that stimulates the transcription of androgen-responsive genes. Constitutive AR expression is restricted to metastatic prostate cancer (Linja & Visakorpi 2004).

2.2.2 Tumour suppressor genes

Tumour suppressor genes, also called anti-oncogenes, function as gatekeepers that negatively regulate normal cell growth. They are involved in the inhibition of cell proliferation, regulation of the cell cycle and apoptosis, cell adhesion and transcriptional regulation. The loss of these genes leads to uncontrolled cell division and growth. Mutations that inactivate tumour suppressor genes are usually recessive because they lead to loss of function (Levine 1990). The most commonly observed inactivating changes include point mutations (often nonsense or frameshift mutations), deletions, chromosomal rearrangements and methylation of promoter regions, all of which lead to loss of heterozygosity (LOH). According to Knudson's classic two-hit hypothesis, in hereditary cancer, LOH is inherited due to a germline mutation, whereas in sporadic cancer, both inactivating mutations occur in tumour tissue (Knudson 1971).

One of the most critical tumour suppressors in prostate cancer is *PTEN*, the *Phosphatase And Tensin Homolog* gene, which is frequently mutated in a large variety of human cancers. PTEN phosphatase deactivates phosphoinositide-3-kinase (PI3K)-dependent signalling which influences cell proliferation, survival and invasion (Barbieri et al. 2013). The *PTEN* locus at 10q23 is deleted in approximately 40% of primary prostate cancers and inactivated in 5-10% of advanced cancers (Cairns et al. 1997, Barbieri et al. 2013). Another gene that is commonly inactivated in epithelial cancers is *RB1 (Retinoblastoma 1)* at 13q14, the first tumour suppressor gene to be identified (Knudson 1971). Under normal conditions, RB prevents cells from entering into the cell cycle and cell division. In cancer, RB regulation is lost due to mutation or deletion, which leads to aberrant cell proliferation (Burkhart & Sage 2008). The inactivation of *RB1* is a rare event in localized prostate cancer but has been detected in approximately 45% of advanced, incurable cancers (Sharma et al. 2010). In addition, mutations and deletions abolishing the tumour protein p53 (TP53) function have been observed in up to 40% of prostate cancers (Barbieri et al. 2013, Khemlina et al. 2015). *TP53* encodes a sequence-specific TF responsible for maintaining genomic stability. Under cellular stress, p53 activates the transcription of genes involved in cell cycle arrest, apoptosis, senescence and DNA repair.

2.2.3 DNA repair genes

When cells duplicate their DNA before cell division, errors occasionally occur. DNA repair genes code for proteins responsible for correcting these replication errors. The main function of DNA repair genes is to maintain genome stability by restoring the correct nucleotide sequence. The inactivation of these genes leads to failure in repair, which results in the accumulation of additional mutations in the cell. This genomic instability likely contributes to neoplastic transformation (Umar & Kunkel 1996). DNA repair genes are often classified as tumour suppressor genes because both are inactivated by recessive mutations.

The role of DNA repair genes in prostate cancer is minor. Typically, genetic aberrations are observed in fewer than 10% of patients (Khemlina et al. 2015). The most frequently mutated DNA repair gene is *BRCA2 (Breast Cancer 2, Early Onset)*. Carriers of germline *BRCA2* mutations are at a five-fold higher risk of developing prostate cancer than are non-carriers. *BRCA2* mutations have also been reported to predispose men to more aggressive disease with worse prognosis (Eeles et al. 2014). Somatic mutations in the *ATM (Ataxia Telangiectasia Mutated)* gene have been observed in approximately 5% of prostate cancers. ATM functions as a master controller of cell cycle checkpoint signalling required for DNA damage response (Khemlina et al. 2015). Other DNA repair genes that are occasionally mutated in prostate cancer patients include *CHEK2 (Checkpoint Kinase 2), BRIP1 (BRCA1-Associated C-Terminal Helicase 1), PALB2 (Partner And Localizer of BRCA2), BRCA1 (Breast Cancer 1, Early Onset)* and *PMS2 (Postmeiotic Segregation Increased 2)*. Although rare, mutations in these genes have been suggested to correlate with advanced disease and may therefore prove to be useful in the clinical setting (Leongamornlert et al. 2014).

2.2.4 Epigenetic alterations

Epigenetic alterations are defined as inherited changes in gene expression that do not affect the primary DNA sequence (Strand et al. 2014). They refer to the addition or removal of chemical groups or moieties to DNA or histone proteins, accomplished by enzymes such as DNA methyltransferases, histone methyltransferases or histone acetyltransferases. In normal cells, epigenetic alterations control tissue- and developmental stage-specific gene expression, the silencing of the inactive X chromosome in females, and imprinting, the silencing of individual alleles based on their parental origin. In cancer, these regulatory patterns disintegrate, leading to the aberrant function of hundreds of genes (Weichenhan & Plass 2013).

DNA methylation is a mechanism responsible for long-term gene silencing. This is achieved by the methylation of cytosine residues at CpG islands, repeated CpG dinucleotide regions found in gene promoters. Normally, promoters are unmethylated, allowing active transcription. Promoter hypermethylation is a frequently observed phenomenon in tumour cells. Methylated promoters prevent TFs from binding, thus leading to the inactivation of tumour suppressor genes (Strand et al. 2014). Another alteration that is characteristic of cancer is global hypomethylation, the loss of methylation in intergenic regions and repetitive elements, which may result in the accumulation of chromosomal breaks and rearrangements (Dobosy et al. 2007). Aberrant DNA methylation patterns have been reported in precursor lesions of prostate cancer, such as PIN, in early tumourigenesis and in metastatic cancers, suggesting that epigenetic alterations play a major role in prostate cancer initiation and progression (Damaschke et al. 2013, Strand et al. 2014).

In addition to DNA methylation, transcription is regulated by histone modifications and chromatin structure remodelling (Damaschke et al. 2013). The highly conserved core histone proteins (H2A, H2B, H3 and H4) can be modified by the addition or removal of acetyl, methyl or ubiquitin groups. Generally, acetylation creates an open chromatin structure and is associated with active transcription, whereas deacetylation results in transcriptional repression (Dobosy et al. 2007). The enzymes responsible for the removal of acetyl groups, histone deacetylases (HDACs), are up-regulated in prostate cancer and have been suggested to function as transcriptional co-repressors (Patra et al. 2001). Histones can also be modified by methylation, which affects chromatin conformation and leads to gene silencing. A well-characterized histone methyltransferase, EZH2 (enhancer of zeste homolog 2), is overexpressed in prostate cancer and has been shown to associate with aggressive, metastatic disease (Varambally et al. 2002).

Epigenetics is a field of intensive research, and an increasing amount of knowledge on the disturbed patterns of gene regulation in cancer is beginning to emerge. Understanding the function of the epigenome will undoubtedly aid in understanding the complex molecular mechanisms that drive neoplastic processes within the cell. In the future, information on epigenetic alterations may potentially be used to identify individuals at risk of developing prostate cancer or to design treatment strategies (Damaschke et al. 2013).

2.3 The genetics of inherited prostate cancer risk

Due to genetic and phenotypic heterogeneity, the identification of prostate cancer susceptibility genes and of variants associated with an increased cancer risk has been challenging. What has become evident, however, is that a large number of genes and variants are involved, each with varying penetrance. Efforts aimed at mapping prostate cancer risk loci have predominantly focused on the identification of either rare, highly penetrant variants in prostate cancer families or common, low-risk variants linked to disease risk in the general population (Eeles et al. 2014). While rare variants explain only approximately 5-10% of the overall inherited prostate cancer risk (Demichelis & Stanford 2015), the current estimates of the contribution of common variants are as high as 38.9% (Amin Al Olama et al. 2015). Even so, less than half of the familial risk is currently explained, leaving the majority of the underlying genetic factors unknown.

2.3.1 Candidate genes identified by linkage analysis

The most traditional gene mapping method, linkage analysis, is based on the co-transmission of a genetic marker and disease phenotype in pedigrees. Typically, multiple families with several affected members, their unaffected siblings and their parents are included in the study. The DNA samples of all family members are genotyped for hundreds or thousands of genetic markers, and the inheritance of these markers together with the disorder is then evaluated. If a certain allele of a polymorphic marker is observed in affected family members more often than could be expected by chance, positive linkage between this allele and disease is declared. The strength of linkage is described with LOD (logarithm of odds) score, and LOD scores >3.0 are considered statistically significant (Foulkes 2008). The HLOD (heterogeneity LOD) score is often more useful for complex diseases, where the same phenotype can be caused by mutations in different genes. HLOD combines LOD scores from all analysed sites.

Linkage analysis has proven successful in the identification of genes underlying monogenic Mendelian diseases. In case of complex disorders, the method has been less effective. A few prostate cancer candidate genes have, however, been recognized and are listed in Table 2. Disease-associated variants in these genes are highly penetrant but have a low frequency in the general population (minor allele frequency, MAF ≤1%). Most of the variants are located in protein-coding regions of the genes and, therefore, have a large effect on prostate cancer risk.

Table 2. Prostate cancer candidate genes identified by linkage analysis.

Gene name	Abbreviation	Locus	Reference
Ribonuclease L	*RNASEL*	1q25	Carpten et al. 2002
Macrophage Scavenger Receptor 1	*MSR1*	8p22	Xu et al. 2002
Breast Cancer 2, Early Onset	*BRCA2*	13q12	Edwards et al. 2003
Partner And Localizer of BRCA2	*PALB2*	16p12	Erkko et al. 2007
ElaC Ribonuclease Z 2	*ELAC2*	17p11	Tavtigian et al. 2001
Homeobox B13	*HOXB13*	17q21	Ewing et al. 2012
Checkpoint Kinase 2	*CHEK2*	22q12	Dong et al. 2003, Seppälä et al. 2003a

The three major candidate genes responsible for prostate cancer susceptibility in Finland are *HOXB13, CHEK2* (Seppälä et al. 2003a) and *RNASEL* (Carpten 2002), whereas the role of *ELAC2, MSR1, BRCA2* and *PALB2* is either small or completely non-existent (Rökman et al. 2001, Seppälä et al. 2003b, Ikonen et al. 2003, Pakkanen et al. 2009). Linkage mapping has also been useful in the identification of other genomic loci that are associated with increased prostate cancer risk in Finland, including 3p25-p26, 11q13-q14 (Schleutker et al. 2003) and Xq27-q28 (Xu et al. 1998). Recently, a potential candidate gene located at 11q13.5, *EMSY (C11orf30)* was shown to associate with aggressive prostate cancer and prostate cancer mortality (Nurminen et al. 2013). In contrast, elaborate studies aiming at discovering the causative genes at 3p25-p26 and Xq27-q28 have remained unsuccessful (Kouprina et al. 2005, Rökman et al. 2005, Kouprina et al. 2007, Bailey-Wilson et al. 2012).

2.3.2 Common variants identified by association analysis

Association analysis aims at finding evidence for the co-occurrence of disease phenotype and a certain marker allele or haplotype in the general population. It is based on linkage disequilibrium (LD), the non-random association of alleles. In practice, this means that the alleles at nearby loci are observed together more often than what would be expected by chance. Association studies exploit large population samples and are conducted using the case-control setting. Typically, hundreds of thousands or even millions of markers are genotyped in hundreds or thousands of individuals simultaneously. Allele frequencies are then compared between patients and controls in order to detect alleles that are over-represented among patients and may therefore be involved in disease susceptibility (Spans et al. 2013). These genome-wide association studies (GWAS) are effective in finding common disease alleles.

The alleles identified by GWAS are often located in non-coding regions of the genome (Xu et al. 2014). They have a high frequency in the general population (MAF ≥5%) but show only a weak to modest effect on prostate cancer risk (average OR: 1.1 – 1.3) (Demichelis & Stanford 2015). This is known as the common disease, common variant principle. One of the first studies that applied GWAS in prostate cancer genetics reported a disease-associating variant on 8q24 (Amundadottir et al. 2006). Subsequent analyses have confirmed the association, refined it into three independent regions within 8q24 and verified the importance of this locus in prostate cancer susceptibility (Gudmundsson et al. 2007a, Haiman et al. 2007, Yeager et al. 2007, Jin et al. 2012). Since 2006, a vast number of GWAS and meta-analyses combining the results from individual studies have been performed and numerous prostate-cancer-associated single nucleotide polymorphisms (SNPs) have been published. The findings are listed in a manually curated, quality controlled GWAS Catalog (www.ebi.ac.uk/gwas/), developed in collaboration between the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI) (Welter et al. 2014). Currently, the catalog contains results from 28 GWAS reporting 193 SNPs that associate statistically significantly ($p \leq 1.0$ x 10^{-5}) with prostate cancer (accessed: 26 Nov, 2015). According to the GWAS Catalog, prostate-cancer-associated SNPs have been detected in all chromosomes except for the Y chromosome. Most GWAS hits are located on chromosomes 2, 3, 6, 8, 10, 11, 17 and X. Several novel candidate genes for HPC have been identified by GWAS, including *HNF1B (Hepatocyte Nuclear Factor 1-Beta)* at 17q12 (Gudmundsson et al. 2007b) and *MSMB (Microseminoprotein Beta)* at 10q11.2 (Thomas et al. 2008).

At present, the clinical significance of the common non-coding variants remains largely unknown. However, multiple common variants in the same individual have been shown to increase prostate cancer risk (Zheng et al. 2008, Eeles et al. 2013), especially if the patient has a positive family history of the disease (Lindström et al. 2012). A recent study demonstrated that prostate cancer risk was highest for carriers of 15-16 common, low-risk alleles (OR = 3.0, 95% CI 2.0 – 4.4). In addition, familial patients were observed to carry more risk alleles than were unselected population cases (Teerlink et al. 2014).

2.3.3 Germline copy number variation analysis

Over the last few years, the contribution of unbalanced, structural genomic variants to complex human disorders has been increasingly appreciated. Submicroscopic variants involving the gain or loss of genetic material have been termed copy number variants (CNVs). By definition, a CNV is a DNA segment ranging from 1 kb to 3 Mb in size whose copy number differs from that of the reference genome (Feuk et al. 2006). CNVs can either form *de novo* or be inherited. They result from chromosomal rearrangements, including deletions, duplications, insertions and translocations, and are estimated to comprise as much as 13% of the human genome (Stankiewicz & Lupski 2010). On average, each individual carries approximately 1,300 CNVs with a median size of 2.9 kb (Conrad et al. 2010). An inverse correlation between CNV size and frequency has been observed, and CNVs larger than 100 kb are rare (<1%) in the general population (Itsara et al. 2009). While most CNVs are benign polymorphisms, several variants have been implicated in complex human disorders, ranging from neurological, cardiovascular and metabolic diseases to asthma and cancer (Almal & Padh 2012). CNVs mediate their deleterious phenotypic effects by altering gene dosage, perturbing the regulation of gene expression or disrupting the coding sequence of a gene (Stranger et al. 2007a).

Rare germline CNVs, varying from 10 kb to >100 kb in size, have been suggested to contribute to cancer predisposition, especially in high-risk cancer families (Kuiper et al. 2010). CNVs can promote tumourigenesis by several mechanisms: a tumour suppressor gene or a DNA repair gene can be deleted, an oncogene can be amplified, or a regulatory element can be removed or introduced to a new genomic location, thereby leading to aberrant gene expression (Kuiper et al. 2010, Krepischi et al. 2012a). Indeed, an association between inherited CNVs and increased cancer risk has recently been demonstrated for childhood neuroblastoma (Diskin et al. 2009), colorectal cancer (Venkatachalam et al. 2011), breast cancer (Krepischi et al. 2012b, Kuusisto et al. 2013) and endometrial cancer (Moir-Meyer et al. 2015). The involvement of germline CNVs in prostate cancer susceptibility has also been investigated, and a few statistically significant associations have been identified (Table 3). Two of these loci, 2p24.3 and 20p13, were shown to associate with an aggressive form of the disease (Liu et al. 2009b, Jin et al. 2011).

Table 3. Germline CNVs that associate significantly with increased prostate cancer risk.

Locus	Gene symbol	CNV type	CNV size	Population	Reference
2p24.3	none	Deletion	5.9 kb	Caucasian	Liu et al. 2009b
12q21.31	*MGAT4C*[a]	Deletion	7.0 kb	Caucasian	Demichelis et al. 2012
14q32.33	*IGHG3*[b]	Duplication	9.4 kb	African American	Ledet et al. 2013
15q21.3	none	Deletion	5.7 kb	Caucasian	Demichelis et al. 2012
20p13	*SIRPB1*[c]	Deletion	32.3 kb	Caucasian	Jin et al. 2011

[a] *Mannosyl-Glycoprotein Beta-1,4-N-Acetylglucosaminyltransferase, Isozyme C*

[b] *Immunoglobulin Heavy Constant Gamma 3*

[c] *Signal-Regulatory Protein Beta 1*

In contrast to inherited, germline changes, the number of somatic CNVs detected in prostate tumours is remarkably high. A recent meta-analysis combined the results from eleven CNV studies performed on 662 primary or advanced prostate tumours and reported 14 recurrent deletions and five recurrent gains in the dataset. The most frequent somatic copy number change was the deletion of chromosome 8p. Other common CNVs included gain of 8q, losses at 2q, 3p, 5q, 6q, 13q, 16q, 17p and 18q, deletions involving the *PTEN* gene, and *TMPRSS2-ERG* gene fusions (Williams et al. 2014).

2.4 Prostate cancer susceptibility loci at 2q37 and 17q11.2-q22

A positive linkage signal on the long arm of chromosome 2 was detected for the first time in a genome-wide screening of 504 North-American brothers with prostate cancer (Suarez et al. 2000). The screening, which was based on 420 highly polymorphic microsatellite markers, identified a large region extending from 2q32.1 to 2q37.3. The signal was subsequently confined to 2q37.2-q37.3 in a study focusing on 12 American HPC families with the co-occurrence of pancreatic cancer (Pierce et al. 2007). Suggestive evidence for linkage on chromosome 17q (LOD = 2.36) was obtained in a genome-wide linkage scan of 175 predominantly Caucasian families participating in the University of Michigan Prostate Cancer Genetics Project (Lange et al. 2003). The results from an extended study, involving additional pedigrees from Johns Hopkins University and exploiting a total of 24 microsatellite markers on chromosome 17, were reported a few years later (Lange et al. 2007). This refined analysis narrowed the linkage peak to 17q21-q22, thus confirming the linkage signals that had been observed for this region in two previous, combined genome-wide linkage scans of 426 and 1233 HPC families (Gillanders et al. 2004, Xu et al. 2005). Further evidence for the importance of the chromosomal regions 2q37 and 17q12-q24 in prostate cancer susceptibility has subsequently been obtained in several GWAS (e.g., Gudmundsson et al. 2007b, Eeles et al. 2008, Kote-Jarai et al. 2011, Schumacher et al. 2011, Jin et al. 2012).

The interconnection of these two loci of interest with increased prostate cancer risk in Finland was examined in a genome-wide linkage scan performed on 69 Finnish high-risk HPC families (Cropp et al. 2011). Altogether, 413 microsatellite markers and 6008 SNPs were genotyped, and genotype data were combined with phenotype and pedigree information. Significant linkage peaks with HLOD scores >3.3 were observed in chromosomal regions 2q37.3 and 17q12-q21.3 (Figure 2), further confirming the association of these two loci with hereditary prostate cancer (Cropp et al. 2011). Several prostate-cancer-associated genes reside in these two loci, including the known risk gene *HOXB13* and the candidate genes *ZNF652, HDAC4* and *ANO7*. All four genes are expressed in the prostate and, except for *ANO7*, are involved in transcriptional regulation. The *ANO7*-encoded membrane protein likely participates in cell-cell interactions on the prostate epithelium.

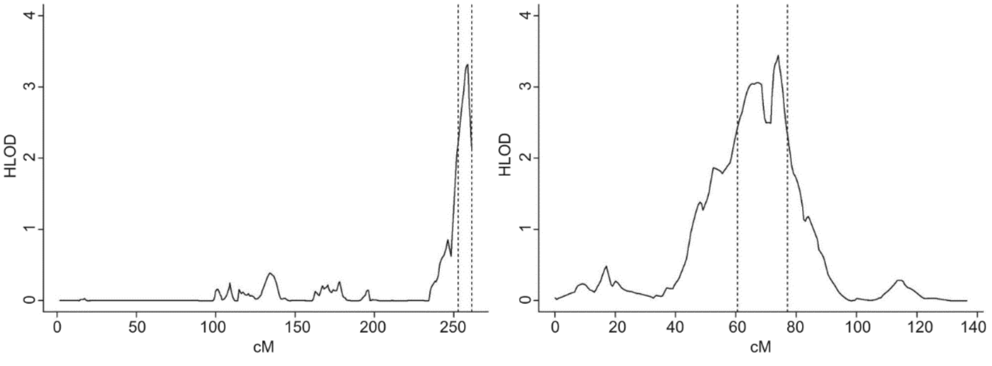


Figure 2. Individual HLOD plots for chromosomes 2 (left) and 17 (right) from the linkage analysis results for 69 Finnish prostate cancer families using the combined SNPs and microsatellite data. The HLOD linkage results are 3.32 for chromosome 2 and 3.44 for chromosome 17. cM denotes CentiMorgan, a genetic linkage unit that corresponds to approximately 1 Mb. (Adapted from Cropp et al. 2011). Reprinted with permission from John Wiley and Sons.

2.4.1 HOXB13

The *HOX (Homeobox)* genes are critical developmental genes. They encode TFs that regulate key pathways during vertebrate embryogenesis and are responsible for proper anterior-posterior pattern formation (Bhatlekar et al. 2014). The human genome contains 39 *HOX* genes distributed into four separate gene clusters (A-D) on chromosomes 7p14, 17q21, 12q13 and 2q31, respectively (Quinonez & Innis 2014). In addition to controlling the normal development of various tissues, *HOX* genes have been found to be involved in the development of several cancers, such as breast and ovarian cancer, colon cancer, prostate cancer and lung cancer (reviewed in Bhatlekar et al. 2014). In tumours, *HOX* genes are either up- or down-regulated. These aberrant expression patterns indicate that *HOX* genes play a central role in maintaining normal adult tissue homeostasis.

The *Homeobox B13 (HOXB13)* gene belongs to the evolutionary conserved *HOXB* gene cluster located on chromosome 17q21. HOXB13 is essential for prostate organogenesis (Huang et al. 2007a) and is highly expressed in normal prostate cells and in prostate cancer cells. HOXB13 has been shown to repress the expression of androgen-responsive genes by interacting with the androgen receptor (Norris et al. 2009). In androgen-independent tumours, the high overexpression of HOXB13 has been reported to be associated with the growth advantage of prostate cancer cells (Kim et al. 2010). A recent study described the functional interaction between HOXB13 and a prostate cancer susceptibility variant, rs339331 at 6q22. The T allele of rs339331 was observed to enhance the binding of HOXB13 to a transcriptional enhancer, thereby leading to the allele-specific up-regulation of the *Regulatory Factor X 6 (RFX6)* gene. In prostate cancer, increased *RFX6* expression has been shown to associate with tumour progression, metastasis and risk of biochemical relapse (Huang et al. 2014). HOXB13 has also been demonstrated to enhance the invasive potential of prostate cancer cells, predominantly by down-regulating the expression of prostate-epithelium-specific ETS transcription factor, PDEF (Kim et al. 2014).

The association between the *HOXB13* gene and prostate cancer risk was first described in 2012 (Ewing et al. 2012). In this study, 202 genes in the 17q21-q22 region were screened for germline variants using DNA samples from 94 unrelated HPC patients. Index cases from four families were found to be heterozygous for the c.251G>A, p.G84E variant (rs138213197) in the *HOXB13* gene. Testing of additional affected and unaffected family members, unselected prostate cancer patients and control subjects revealed that the p.G84E variant co-segregated with prostate cancer, especially in patients of European origin. Furthermore, the variant was observed to associate statistically significantly with early-onset familial disease. These results have subsequently been replicated in a number of studies (e.g., Akbari et al. 2012, Breyer et al. 2012, Karlsson et al. 2014, Xu et al. 2013).

2.4.2 *ZNF652*

Approximately 3% of the human genome consists of genes coding for zinc finger proteins, which regulate a vast variety of biological processes (Klug 2010). The diverse functions of zinc finger proteins include DNA recognition, RNA packaging, transcriptional activation and repression, regulation of apoptosis, protein folding and assembly, and lipid binding (Laity et al. 2001). Zinc finger is a structural motif, a folded protein domain that is stabilized by a zinc ion. The classical and most abundant zinc-binding motif contains two cysteines and two histidines ligated to a zinc ion and is known as the Cys_2His_2 or C_2H_2 motif (Laity et al. 2001). Several zinc fingers can be linked in tandem and are used to specifically recognize and bind target DNA sequences (Klug 2010).

The *ZNF652* gene, located at 17q21.3, encodes Zinc Finger Protein 652. It contains seven C_2H_2 zinc finger motifs and functions as a DNA-binding transcriptional repressor (Kumar et al. 2006). *ZNF652* is ubiquitously expressed. While its highest expression levels have been observed in normal breast, vulva, prostate and pancreas cells, its expression is generally down-regulated in primary tumours of the corresponding tissues (Kumar et al. 2006). However, approximately half of the prostate tumours have been reported to maintain high levels of both ZNF652 and AR expression, which predispose patients to an increased risk of PSA relapse (Callen et al. 2010). ZNF652 has also been shown to form a complex with CBFA2T3 (Core-Binding Factor, Alpha Subunit 2, Translocated to, 3) (Kumar et al. 2006). CBFA2T3, a candidate breast cancer tumour suppressor (Kochetkova et al. 2002), enhances the repressor activity of ZNF652 (Kumar et al. 2006).

The identification of the ZNF652 consensus DNA-binding sequence (Kumar et al. 2008) led to the discovery of 113 ZNF652 target genes, many of which have been linked to various cancers, including prostate cancer (Kumar et al. 2011). So far, only two *ZNF652* variants have been reported to associate with increased prostate cancer risk. Rs7210100, described with a frequency of 4-7% in African-American men, is rare (<1%) in non-African populations (Haiman et al. 2011). Another SNP, rs11650494, located downstream of the *ZNF652* gene, has been suggested to represent a European-specific risk variant (Eeles et al. 2013).

2.4.3 *HDAC4*

Histone deacetylases (HDACs) are ubiquitously expressed transcriptional repressors that play an important role in the regulation of transcription, the progression of the cell cycle and various developmental events. Instead of directly binding to DNA, HDACs exert their function by removing acetyl groups from lysine residues in the core histones, thereby inducing a conformational change in chromatin structure. The tightly condensed chromatin then prevents transcriptional co-factors from accessing their binding sites (Stelzer et al. 2011). In mammals, at least 18 HDACs have been identified and can be divided into three classes (I-III) based on their size, sequence homology and catalytic properties (Wang et al. 2014). Class II HDACs are able to shuttle between the nucleus and cytoplasm (Fischle et al. 2001). One of the key members of class II HDACs is HDAC4, encoded by the *Histone Deacetylase 4 (HDAC4)* gene at 2q37.3. It is widely expressed in a variety of tissues and regulates its target genes by interacting with the Myocyte Enhancer Factor 2 (MEF2) family of TFs (Wang et al. 1999).

Homozygous deletions of *HDAC4* have been observed in a genome-wide copy number analysis of 76 melanoma cell lines, suggesting tumour suppressor activity for this gene (Stark & Hayward 2007). An *in silico* genomic pathway analysis revealed that *HDAC4* activation correlated with inflammatory processes in breast cancer and glioblastoma cells (Cohen et al. 2013). In androgen-independent prostate adenocarcinomas, HDAC4 has been observed to accumulate in the nucleus. It has been proposed that HDAC4 contributes to the development of aggressive, hormone-refractory prostate cancer by suppressing genes that induce differentiation (Halkidou et al. 2004). HDAC4 has also been reported to repress HOXB13 expression in co-operation with the ubiquitous transcription factor YY1 (Yin And Yang 1 Protein) (Ren et al. 2009). In addition to deacetylation, HDAC4 can repress transcriptional activity by another mechanism, SUMOylation, which refers to the enzymatic addition of a small ubiquitin-like modifier (SUMO) to a substrate protein (Chen & Lu 2015). HDAC4 has been demonstrated to repress AR expression in prostate cancer cells through the SUMOylation of the endogenous AR (Yang et al. 2011).

2.4.4 *ANO7*

Anoctamin 7 (*ANO7*) belongs to the *TMEM16* family of genes encoding membrane proteins. It was described more than a decade ago as a gene expressed only in normal prostate and prostate cancer cells (Bera et al. 2004). *ANO7*, located at 2q37.3, has several aliases, including *NGEP* for *New Gene Expressed in Prostate*, *TMEM16G* for *Transmembrane Protein 16G* and *D-TMPP* for *Dresden-Transmembrane Protein of the Prostate*. Characterization of the ANO7 protein structure revealed that it has eight transmembrane domains and intracellular N- and C-termini (Das et al. 2008). ANO7 has been proposed to function as a phospholipid scramblase, an enzyme embedded in the plasma membrane that transports phospholipids between the two lipid layers (Picollo et al. 2015). Phospholipid scramblases have been suggested to participate in disrupting the cell membrane in response to cell activation, injury or apoptosis (Sahu et al. 2007). ANO7 may also function as an ion channel or a co-regulator of other ion channels. However, at present, the role of ANO7 is poorly understood and requires further research (Picollo et al. 2015).

Due to alternative splicing, long and short transcript variants of the *ANO7* messenger RNA (mRNA) exist, encoding long and short isoforms of the ANO7 protein, respectively. The long isoform is transported to the plasma membrane (Bera et al. 2004) and has been suggested to play a role in cell-cell interactions and prostate cell adhesion (Das et al. 2007). The short isoform is intracellular and is possibly located on the endoplasmic reticulum (Duran et al. 2012). A recent tissue microarray analysis showed that ANO7 expression is down-regulated in prostate cancer. In addition, an inverse correlation between the level of ANO7 expression and the degree of malignancy was found (Mohsenzadegan et al. 2013). *ANO7* variants associated with prostate cancer have not yet been identified. However, an analysis of 98 breast cancer exomes revealed that *ANO7* mutations may influence breast cancer development and prognosis (Li et al. 2015). In this study, the frequency of deleterious *ANO7* mutations was observed to increase with tumour malignancy.

2.5 Next-generation sequencing technologies

The first full human genome sequence was published in April 2003 as a result of a collaborative, international research program, the Human Genome Project (www.genome.gov/10001772). Sequencing was based on the traditional Sanger sequencing method, coupled with automated DNA sequencers. In total, the project cost approximately 300 million US dollars and lasted 13 years (Metzker 2010). During the past decade, scientific discoveries have led to the development of high-throughput sequencing technologies, often referred to as next-generation sequencing (NGS) or massively parallel sequencing. These novel platforms enable the simultaneous analysis of multiple genes of interest at substantially lower costs. Currently, the sequencing of the entire human genome can be accomplished in hours and, according to NHGRI's Genome Sequencing Program, the average cost is 1,363 dollars (www.genome.gov/sequencingcosts). The costs are expected to decrease even further as Illumina (San Diego, CA, USA), a leader in the DNA sequencing industry, has recently presented the new HiSeq X Ten Sequencing System, which is able to sequence 18,000 human genomes per year (49 genomes per day) at the price of 1,000 dollars per genome (www.illumina.com/systems/hiseq-X-sequencing-system/).

2.5.1 Key principles of NGS

The basic NGS workflow is illustrated in Figure 3. It consists of three steps: template generation, sequencing reactions and detection, and data analysis (Rizzo & Buck 2012). First, a library of sequencing reaction templates is prepared. The starting material, usually double-stranded DNA, is fragmented into small sizes, typically ranging from 200 bp to 250 bp (Metzker 2010). Fragments of desired length are then selected for adapter ligation. Adapters are needed in the subsequent target amplification and sequencing steps. Most NGS platforms exploit the sequencing-by-synthesis (SBS) principle, whereby the sequence of the template strand is obtained during the enzymatic synthesis of the complementary strand (Mardis 2008). The detection of incorporated nucleotides is commonly based on optical methods visualizing fluorescent labels, a strategy used in Illumina's MiSeq and HiSeq sequencers (Metzker et al. 2010). Ion Torrent™ applies semiconductor-based sequencing technology (Thermo Fisher Scientific, Waltham, MA, USA). In this SBS approach, the incorporation of nucleotides is visualized as a change in pH, resulting from the release of hydrogen ions during phosphodiester bond formation. Recently, a novel Nanopore sequencing technology was introduced (Oxford Nanopore Technologies, Oxford, UK). Here, an electric current is applied across a protein nanopore. The transportation of single-stranded nucleic acids through the nanopore modulates the electric field, and the change is characteristic for each nucleotide (Luthra et al. 2015).

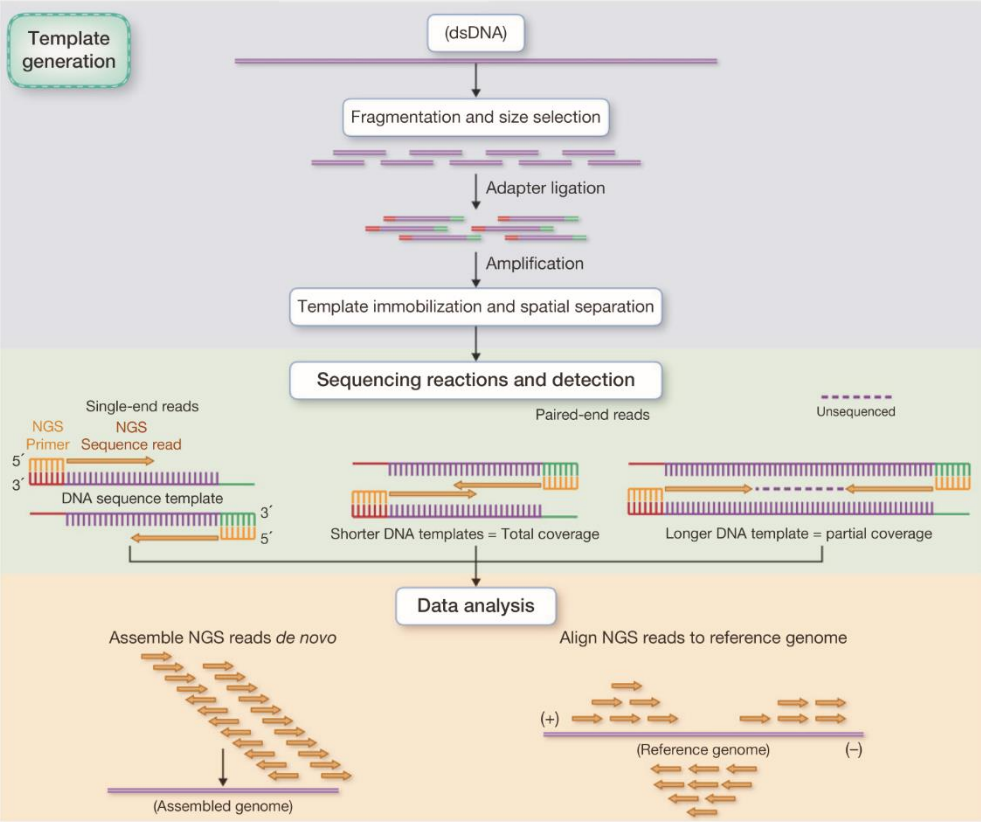

Template generation
(dsDNA)
Fragmentation and size selection
Adapter ligation
Amplification
Template immobilization and spatial separation
Sequencing reactions and detection
Single-end reads
NGS Primer
NGS Sequence read
5´
3´
DNA sequence template
3´
5´
Paired-end reads
Unsequenced
Shorter DNA templates = Total coverage
Longer DNA template = partial coverage
Data analysis
Assemble NGS reads *de novo*
(Assembled genome)
Align NGS reads to reference genome
(+)
(Reference genome)
(–)

Figure 3. Basic workflow for NGS experiments. Sequencing templates are generated from double-stranded DNA (dsDNA), which is fragmented, amplified and sequenced. Data analysis refers to genome assembly, and in human studies, reference genomes are always used. (Adapted from Rizzo & Buck 2012). Reprinted with permission from Michael J. Buck and the American Association for Cancer Research.

NGS data analysis begins with base-calling, the translation of the sequencing signal into base sequences. Then, sequence reads are either assembled *de novo* (built from scratch) or aligned to a reference sequence or genome. The next step, variant calling, aims at identifying genomic alterations by comparing the correctly targeted reads to their reference sequence (Rizzo & Buck 2012). Different variant calling algorithms may be required to reliably recognize different types of variations, but in principle, NGS methods are able to detect single-nucleotide variants, small and large insertions and deletions, copy number variants and gene fusions simultaneously (Luthra et al. 2015).

NGS methods typically provide tens or hundreds of sequence reads representing each target region, which increase the sensitivity and reliability of mutation detection. Sequencing coverage or depth describes the average number of times a base pair has been sequenced. To overcome biases resulting from sequencing errors and uneven read distribution across the reference sequence, a coverage of approximately 30x to 40x is recommended for the accurate identification of variants (Lohmann & Klein 2014). The detection of large genomic rearrangements, repetitive sequences, gene fusions and novel transcripts can be further improved by paired-end sequencing. In this approach, the DNA fragment is sequenced from both ends using the adapters ligated in the template generation step as sequencing primers. Paired-end sequencing is routinely applied in current NGS projects because it also increases the accuracy of alignment, thereby improving the quality of the entire dataset (Rizzo & Buck 2012, Luthra et al. 2015).

2.5.2 NGS applications

Whole-genome sequencing (WGS) refers to the re-sequencing of the entire genome of a cell, the determination of the sequence of all 3 billion base pairs. In addition to protein-coding genes, intergenic and regulatory regions will also be covered (Barbieri et al. 2013). In cancer research, WGS enables the identification of novel disease-associated genetic aberrations, such as gene fusions and balanced chromosomal rearrangements, which are difficult or impossible to identify with traditional mutation detection methods (Rizzo & Buck 2012). A recent whole-genome paired-end sequencing performed on a primary prostate cancer patient and a prostate cancer cell line discovered a total of 21 novel fusion transcripts with functional consequences (Teles Alves et al. 2015). One of the major drawbacks of WGS is the fact that approximately 99% of sequencing data represent the non-coding part of the genome, the function of which is rather poorly characterized. This limits the interpretation, practical usefulness and cost-efficiency of WGS data (Barbieri et al. 2013). Another challenge of WGS and other NGS applications is the management and storage of the vast amount of sequencing data that are generated, typically several hundreds of gigabases per sequencing run (Luthra et al. 2015). Current guidelines recommend the storage of files required to repeat the whole-genome analysis (Aziz et al. 2015).

Whole-exome sequencing (WES) focuses on the sequencing of protein-coding regions only. These represent approximately 1% of the genome (ENCODE Project Consortium 2012). Compared to WGS, WES is a cost-effective and highly sensitive mutation detection approach, as it covers only a limited region of the genome (Barbieri et al. 2013). According to Spans and colleagues, more than 200 prostate-cancer-related exome sequencing reports have already been published. Most of these studies have investigated the different stages of prostate cancer tumourigenesis as well as the progression of the disease to CRPC by sequencing tumour cell exomes (reviewed in Spans et al. 2013). WES has also been used for the exploration of genetic predisposition to prostate cancer. A novel susceptibility gene, *BTNL2 (Butyrophilin-like 2)*, was identified by the exome sequencing of hereditary prostate cancer families (Fitzgerald et al. 2013). Another WES project led to the discovery of 43 nonsense and missense variants associated with familial prostate cancer (Johnson et al. 2014).

Additional, fine-tuned NGS applications include sequencing the coding regions of the approximately 3,000 known disease genes (the "Mendelianome") and targeted gene panels consisting of gene sets relevant to the disease under study (Rizzo & Buck 2012). Numerous predesigned and custom-made gene panels are commercially available and widely used in clinical laboratories (Luthra et al. 2015). It is also possible to re-sequence any DNA region of interest. This approach is known as targeted re-sequencing and requires the selective enrichment of genomic target regions prior to sequencing. The selection of the enrichment method depends on sample type (fresh, frozen or formalin-fixed and paraffin-embedded, FFPE, samples) and the quantity and quality of DNA or RNA. The most commonly used target enrichment strategies include PCR-based enrichment and probe-hybridization-based capture technologies (Luthra et al. 2015). An advantage of targeted sequencing strategies is that they provide higher coverage, which results in an increased accuracy of mutation detection (Rizzo & Buck 2012).

RNA sequencing (RNA-seq), also known as whole-transcriptome sequencing, can be used to sequence all RNA molecules within a cell, including mRNAs, microRNAs (miRNAs) and other non-coding RNAs (Mardis & Wilson 2009). The transcriptome refers to all of the DNA sequences that are transcribed into RNA. Before sequencing, RNA molecules need to be converted to complementary DNA (cDNA) by reverse transcription (Pickrell et al. 2010). RNA-seq provides quantitative information on mRNA expression levels and can be used to investigate expression profiles among different cells or tissues. RNA-seq data also enable the detection of allele-specific expression, the verification of the effect of nonsense mutations, and the identification of alternatively spliced isoforms or fusion transcripts (Mardis & Wilson 2009). Modifications to the standard RNA-seq method allow the mapping of transcription start sites, the identification of antisense transcripts by strand-specific sequencing and small RNA profiling (Ozsolak & Milos 2011). Recently, RNA-seq has been proven to be the method of choice in expression quantitative trait loci (eQTL) analysis (Majewski & Pastinen 2011, Lappalainen et al. 2013, Larson et al. 2015).

2.6 Expression quantitative trait loci (eQTL) analysis

The majority of cancer-associated genetic variants identified by GWAS have been localized in non-coding regions of the genome, such as intronic, intergenic and gene desert regions (Xu et al. 2014). These loci are likely to contain regulatory elements that control the expression levels and patterns of nearby genes. Interestingly, transcriptional regulation has been shown to play an important role in cancer predisposition (Monteiro & Freedman 2013). One fundamental method for studying the function of regulatory regions is the eQTL analysis (Michaelson et al. 2009). eQTLs are defined as genomic regions containing DNA sequence variants that control the expression of one or more genes (Veyrieras et al. 2008). These regulatory elements have been shown to be highly heritable (Wright et al. 2014) and explain differences in gene expression levels among individuals and populations (Nica & Dermitzakis 2013). eQTLs can modify the disease phenotype by affecting the penetrance of rare deleterious variants (Lappalainen et al. 2011).

Standard eQTL analysis estimates the association of SNP genotypes with gene expression levels across tens or hundreds of individuals. Regulatory variants are often located in *cis*, near their target genes (Göring et al. 2007). Therefore, SNPs located within a predefined window, typically within 1 Mb of either side of the target gene, are selected for the eQTL analysis. The correlation between the SNP genotype and the expression level of the target gene is then evaluated. In the traditional approach, gene expression data were obtained using expression microarrays, but currently, transcriptome quantification is largely performed by RNA sequencing (Majewski & Pastinen 2011). The principle of eQTL mapping is illustrated in Figure 4. Recent eQTL mapping studies have identified hundreds of *cis*-eQTLs associated with increased prostate cancer risk (e.g., Shan et al. 2013, Siltanen et al. 2013, Li et al. 2014, Xu et al. 2014, Han et al. 2015, Larson et al. 2015). Further functional analyses are required to validate the causality of the reported candidate variants.

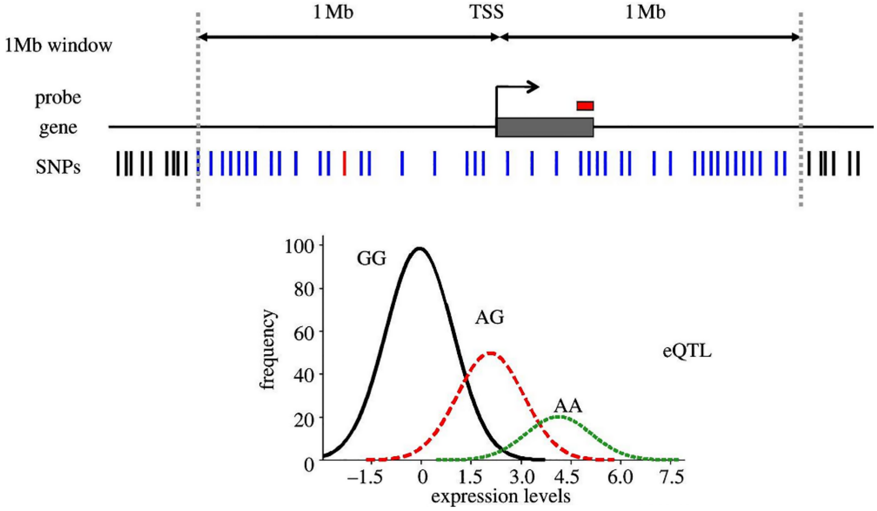


Figure 4. A typical eQTL; many SNPs tested against expression levels as measured by a probe or by other means. The panel below illustrates the difference in the distribution of expression values stratified by the SNP genotype (GG, AG or AA) of the most significant SNP. TSS = transcription start site. (Adapted from Nica & Dermitzakis 2013). Reprinted with permission from The Royal Society.

2.7 Predicting the pathogenicity of novel sequence variants

Due to rapidly evolving sequencing technologies, an increasing number of novel sequence variants are being detected in patients' samples. The assessment of the clinical significance of individual variants can be challenging, especially when they are located in the non-coding regions of the genome. An updated guideline for the classification of sequence variants has recently been published by the American College of Medical Genetics and Genomics (ACMG; Richards et al. 2015). To estimate the pathogenic potential of a variant, several aspects need to be considered, including the allelic frequency, the degree of evolutionary conservation and the effect of the base change on the biochemical properties of the protein. A review of the relevant scientific and medical literature is also strongly recommended (Richards et al. 2015).

2.7.1 Assessing the relevance of the candidate gene

Before evaluating the impact of individual variants on gene function, the biological role of a gene in disease susceptibility needs to be established. Candidate gene selection is often based on the "guilt by association" principle, referring to genes that are either functionally or structurally similar to known disease genes (Patnala et al. 2013). Several different web-based interfaces, tools and knowledge bases have been developed to aid in the evaluation of both the evolutionary conservation and the functional importance of the genes in diverse biological processes. Two frequently used methods to obtain biological information for a given set of genes include gene ontology annotation and pathway enrichment analysis. The Gene Ontology (GO) Project (www.geneontology.org/) classifies genes and their products into three structured ontologies based on biological relationship. These ontologies include the molecular function at the biochemical level, the biological process in which the gene participates and the cellular component in which the gene product is located (Ashburner et al. 2000). Pathway analysis provides information on known metabolic, signalling and regulatory pathways, and it concentrates on identifying interactions between genes. The most commonly used pathway analysis resources are the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa & Goto 2000), WikiPathways (Pico et al. 2008) and Pathway Commons (Cerami et al. 2011).

The GO and pathway data are often accessed via data-mining tools, which incorporate genomic information from several public sources. One such tool is

Ensembl BioMart (www.ensembl.org/biomart/), provided by the Ensembl project, which collects, organizes and stores data from several gene and disease databases as well as from international genome mapping projects (Flicek et al. 2013). Another widely used tool is WebGestalt, the Web-based Gene Set Analysis Toolkit (http://bioinfo.vanderbilt.edu/webgestalt/), which combines functional information with computational analysis (Wang et al. 2013).

2.7.2 Database queries

Several public databases exist that contain information on a growing number of variants that have been identified in the human genome. In principle, population databases offer a general view of genetic variants observed in large populations consisting of both healthy and affected individuals, whereas disease- and gene-specific databases describe variants that have been detected primarily in patients. However, a major drawback of many of the databases is that they contain incorrectly classified variants (Cooper & Shendure 2011). Therefore, confirmation of variant pathogenicity should not rely on database information alone.

2.7.2.1 Population databases

Allele frequency is one means of estimating the pathogenic potential of the variant. Typically, disease-causing variants are rare, with a minor allele frequency of <1% (MacArthur et al. 2014). Population databases (Table 4) provide information on variant frequency in different populations. At present, one of the most commonly used databases is the Exome Aggregation Consortium (ExAC) database, where variant data from more than 60,000 unrelated individuals is stored. Another frequently accessed database is 1000 Genomes, which contains whole-genome sequencing data for more than 2,500 samples, including 100 Finnish genomes (Sudmant et al. 2015). The NCBI's (National Center for Biotechnology Information) SNP database, dbSNP, is a collection of short genetic variations (≤50 bp) for multiple species and is a widely used source for retrieving minor allele frequency (MAF) data. While most dbSNP variants are considered polymorphisms, some may be pathogenic (Richards et al. 2015). The Database of Genomic Variants (DGV) consists of human genomic structural variants, such as CNVs and inversions, which have been identified in healthy control samples only (MacDonald et al. 2014). These variants are at least 50 bp in size.

Table 4. Databases frequently used in the assessment of variant pathogenicity. The first four are population databases, while the remaining five represent disease-specific databases.

Database name	Website
Exome Aggregation Consortium (ExAC)	http://exac.broadinstitute.org/
1000 Genomes	www.1000genomes.org/
dbSNP	www.ncbi.nlm.nih.gov/SNP
Database of Genomic Variants (DGV)	http://dgv.tcag.ca/dgv
Online Mendelian Inheritance in Man (OMIM)®	www.omim.org/
Human Gene Mutation Database (HGMD)®	www.hgmd.org/
Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER)	https://decipher.sanger.ac.uk/
Catalogue of Somatic Mutations in Cancer (COSMIC)	http://cancer.sanger.ac.uk/cosmic/
Dragon Database of Genes Implicated in Prostate Cancer (DDPC)	www.cbrc.kaust.edu.sa/ddpc/

2.7.2.2 Disease databases

Disease-specific databases (Table 4) contain information on variants observed in affected individuals, as well as estimates on variant pathogenicity. Probably the most widely used disease database is the Online Mendelian Inheritance in Man (OMIM®), which comprises referenced descriptions of all known Mendelian disorders and more than 15,000 genes. The Human Gene Mutation Database (HGMD®) contains information on more than 141,000 germline variants in more than 5,700 genes associated with inherited diseases (Stenson et al. 2014). Clinically relevant CNVs are catalogued in DECIPHER, a tool suite designed to assist in the interpretation of genomic variants (Firth et al. 2009). Currently, data on more than 27,000 CNVs in more than 18,000 patients are publicly available in DECIPHER.

COSMIC stores information on the effect of somatic mutations in human cancer. COSMIC contains mutation profiles on 136 known cancer genes with full literature curations, including more than two million coding mutations and more than 60 million gene expression variants (Forbes et al. 2015). Prostate-cancer-associated genes are listed in the Dragon Database of Genes Implicated in Prostate Cancer (DDPC). The involvement of these genes in prostate carcinogenesis has been experimentally verified. DDPC provides integrated data on molecular interactions, pathways, gene ontologies, gene regulation and predicted TF binding sites (Maqungo et al. 2011).

2.7.3 Pathogenicity prediction *in silico*

The deleteriousness of sequence variants can also be assessed using computational *in silico* predictive programs. It is recommended to use more than one predictor because different programs use different algorithms and may therefore provide contradictory results for the same variant (Richards et al. 2015). Generally, *in silico* pathogenicity predictions are based on evolutionary conservation, the biochemical properties of the amino acid, the sequence environment and the effect of the variant on protein structure and function (Cooper & Shendure 2011). Currently, the predictions are restricted to coding variants only, focusing mainly on the functional consequences of either missense or splice site variants (Richards et al. 2015). A selection of frequently used missense predictors is presented in Table 5. In addition to single predictors, integrated systems may be applied. One such example is the machine learning-based method PON-P (Pathogenic-or-Not Pipeline), which merges five predictors to evaluate the effect of missense variants on protein function (Olatubosun et al. 2012).

Table 5. Select commonly used *in silico* pathogenicity predictors for missense variants (modified from Richards et al. 2015).

Name	Website	Reference
MutationTaster	www.mutationtaster.org	Schwarz et al. 2014
PANTHER (Protein Analysis Through Evolutionary Relationships)	pantherdb.org	Mi et al. 2010
PhD-SNP (Predictor of human Deleterious Single Nucleotide Polymorphisms)	snps.biofold.org/phd-snp/phd-snp.html	Capriotti et al. 2006
PolyPhen-2 (Polymorphism Phenotyping v2)	genetics.bwh.harvard.edu/pph2	Adzhubei et al. 2013
PROVEAN (Protein Variation Effect Analyzer)	provean.jcvi.org	Choi et al. 2012
SIFT (Sorting Tolerant From Intolerant)	sift.jcvi.org	Kumar et al. 2009
SNAP (Screening for Non-Acceptable Polymorphisms)	www.rostlab.org/services/SNAP	Bromberg & Rost 2007
SNPs&GO	snps-and-go.biocomp.unibo.it/snps-and-go/	Calabrese et al. 2009

2.7.4 Estimating the impact of regulatory variants

In silico pathogenicity predictors are unable to estimate the effect of non-coding sequence variants. Therefore, additional tools are needed to identify variants with putative regulatory potential and to evaluate their impact on gene expression. One such tool is the RegulomeDB, an approach and a database designed to interpret the functional consequences of non-coding variants (Boyle et al. 2012). RegulomeDB combines computational predictions with experimental data on functional DNA elements from various sources and scores the sequence variants from 1 to 6 according to the likelihood that they possess regulatory potential. The smaller the Regulome score, the more likely the variant has functional effects. The majority of the experimental functional data are obtained from the Encyclopedia of DNA Elements (ENCODE) Project. The purpose of this project was to map all functional elements in the human genome, and for 80% of the DNA sequences, a biochemical task was indeed found (ENCODE Project Consortium 2012). The identified elements include RNA transcribed regions, protein-coding regions, TF binding sites, DNA methylation sites and chromatin structure.

3 Aims of the Study

This study aimed to further elucidate the genetic factors that contribute to inherited prostate cancer risk, especially in Finland.

The specific aims were as follows:

1. To study the role of the *HOXB13* variant p.G84E, a known prostate-cancer-associated risk allele, among Finnish prostate cancer patients (I).

2. To search for novel prostate cancer candidate genes and disease-associated sequence variants from chromosomal regions 2q37 and 17q11.2-q22, which have previously been linked to prostate cancer (II).

3. To identify germline copy number alterations across the whole genome that may increase prostate cancer risk (III).

4 Subjects and Methods

4.1 Human subjects (I-III)

The ethnic ancestry of all of the cancer patients and unaffected control individuals whose samples were included in this research project is Finnish. Table 6 summarizes the number of samples analysed in different studies (I-III).

4.1.1 Familial prostate cancer patients (I-III)

Familial prostate cancer samples have been collected since 1995 by the Laboratory of Cancer Genetics in the University of Tampere and Tampere University Hospital (TAUH). The identification of prostate cancer families started in 1988 and was initially based on questionnaires sent to prostate cancer patients living in the TAUH area. Additional patients and their first-degree relatives were identified using the Finnish Cancer Registry and church parish registries. The sample collection project was also advertised in major Finnish newspapers as well as on the TV and radio, and all practicing urologists in Finland were contacted directly (Schleutker et al. 2000). Currently, the collection contains samples from 375 prostate cancer families, with a total of 583 prostate cancer patients and 1,620 unaffected family members.

The most representative 190 prostate cancer families were selected for this research project. In 150 families, at least three family members had been diagnosed with prostate cancer. The remaining 40 families had only two affected family members, but the patients were either first-degree relatives or one of them had received a cancer diagnosis before the age of 60 years. Study II focused particularly on 37 families that had shown linkage to chromosomal region 2q37 (20 families), chromosomal region 17q11.2-q22 (seven families) or both (ten families) in an earlier study (Cropp et al. 2011). For association analyses in Studies I-III, only the index patient was genotyped. Additional patients and unaffected relatives were subsequently analysed to investigate the co-occurrence of identified genetic variants with the disease phenotype.

The main clinical features were determined for the 190 familial index cases. Their average age at diagnosis was 62.8 years. For one-third of the patients (35.5%), the total serum PSA value at diagnosis ranged from 4.1 to 9.9 ng/ml, whereas PSA levels higher than 10.0 ng/ml were observed in 59% of the patients. The Gleason scores for diagnostic prostate biopsies were ≤6 in 48% of the patients and ≥7 in 29% of the patients. PSA progression indicative of recurrent, advanced disease was detected in 26 index cases (13.7%). Altogether, 147 index patients had died, and prostate-cancer-related death had been reported for 67 patients (35%).

Table 6. Summary of samples used in Studies I-III.

Sample type	Study I	Study II	Study III
Familial prostate cancer patients	190/37[a]	63/188/243/84[b]	105/189/66[e]
Unaffected male family members	28	3/112/15[c]	30/80[f]
Female family members	-	2/92[d]	7/64[f]
Unselected prostate cancer patients	3197	1105	-
Male population controls for prostate cancer	923	923	476
Prostate cancer patients from the screening trial	1184	-	-
Unaffected male controls from the screening trial	4544	-	-
BPH patients with a later diagnosis of prostate cancer	254	-	-
BPH patients with BPH only	262	-	-
Familial breast cancer patients	323	-	-
Unselected breast cancer patients	663	-	-
Female population controls for breast cancer	1449	-	-
Familial colorectal cancer patients	57	-	-
Unselected colorectal cancer patients	385	-	-
Male population controls for colorectal cancer	459	-	-
Prostate cancer cell lines	8	-	-
Normal prostate epithelial cell lines	2	-	-
Xenografts	19	-	-

[a] p.G84E genotyping/p.G84E co-segregation analysis

[b] targeted re-sequencing/Sequenom validation (58 variants)/co-segregation analysis (4 SNPs)/RNA-seq

[c] targeted re-sequencing/co-segregation analysis (4 SNPs)/RNA-seq

[d] targeted re-sequencing/co-segregation analysis (4 SNPs)

[e] genome-wide SNP array/CNV validation/co-segregation analysis (*EPHA3* deletion)

[f] genome-wide SNP array/co-segregation analysis (*EPHA3* deletion)

4.1.2 Unselected prostate cancer patients (I, II)

The collection of samples from prostate cancer patients with no known family history of the disease began in 1996. This unselected, population-based sample collection was performed by the Department of Urology in TAUH and was restricted to patients living in the Pirkanmaa area. Clinical data were obtained from hospital records. If additional family members were later diagnosed with prostate cancer, all patient samples from the respective family were transferred from the unselected sample group to the familial sample group. To date, 7,184 samples have been collected.

4.1.3 Screening trial patients (I)

Another group of unselected prostate cancer patients was obtained from the Finnish arm of the European Randomized Study of Screening for Prostate Cancer (ERSPC), which began in the early 1990s and ended in 2006. The purpose of this study was to investigate whether PSA-based population screening could decrease prostate cancer mortality. In Finland, the study subjects were identified from population registries, and men at the ages of 55, 59, 63 and 67 years were recruited. They were randomly assigned to either the screening group or the control group, and their screening continued until the age of 71 years (Schröder et al. 2009).

4.1.4 Breast cancer patients (I)

Familial breast cancer patients belonged to well-defined high-risk breast cancer families. They either had been diagnosed at an early age or had at least three first- or second-degree relatives with breast or ovarian cancer. Blood samples of 86 index patients were collected by the TAUH Genetics Outpatient Clinic between 1997 and 2008 (Kuusisto et al. 2011). Samples of an additional 237 index patients were collected by the Helsinki University Central Hospital (HUCH) Departments of Oncology and Clinical Genetics between 1985 and 1994 (Eerola et al. 2000). Finnish *BRCA1* and *BRCA2* founder mutations were excluded from each familial index case.

The collection of unselected breast cancer samples was organized simultaneously with the familial breast cancer sample collection in HUCH (Eerola et al. 2000). In TAUH, unselected breast cancer samples were collected between 1997 and 1999 (Syrjäkoski et al. 2000). The Helsinki subgroup of unselected breast cancer patients in Study I consisted of 253 samples, and the Tampere subgroup included 410 samples.

4.1.5 Colorectal cancer patients (I)

The population-based collection of colorectal adenocarcinoma samples was organized at nine large regional hospitals in southeastern Finland between 1994 and 1998 (Aaltonen et al. 1998, Salovaara et al. 2000). Fresh-frozen specimens were examined histologically to ensure that the proportion of tumour cells was as high as possible, preferably more than 50%. Official population registries were used to document family histories, and patients classified as familial had at least one first-degree relative with colorectal carcinoma. Cancer diagnoses were verified from the Finnish Cancer Registry and from the Finnish Hereditary Non-Polyposis Colorectal Cancer (HNPCC) Registry.

4.1.6 Patients with benign prostatic hyperplasia (I)

The collection of benign prostatic hyperplasia (BPH) samples was performed by the Urology Outpatient Department of TAUH between 1995 and 2004. The primary reason for the first prostate biopsy was either elevated PSA level (>4.0 ng/ml) or abnormal digital rectal examination. Of the 516 BPH patients who were included, 262 were diagnosed with histologically confirmed BPH only. The remaining 254 patients developed prostate cancer more than one year after the original BPH diagnosis (Saaristo et al. *unpublished results*). Prostate cancer diagnoses were verified from the Finnish Cancer Registry.

4.1.7 Unaffected control individuals (I-III)

Population controls for prostate cancer and colorectal cancer consisted of voluntary male blood donors. Female population controls for breast cancer included 900 voluntary female blood donors in the Tampere subgroup and 549 donors in the Helsinki subgroup. All of the blood donors were anonymous individuals aged between 18 and 65 years and healthy at the time of blood draw. Their blood samples were collected by the Finnish Red Cross Blood Transfusion Service.

Another set of population controls for prostate cancer was obtained from the ERSPC trial. These screening trial controls were unaffected men with a very low total PSA level (<1.0 ng/ml).

In addition, various subsets of unaffected male and female family members belonging to the 190 prostate cancer families were included in all three studies, mainly to investigate the co-occurrence of the identified variants with disease phenotype.

4.1.8 Ethical aspects (I-III)

This research project followed the guidelines of Responsible Conduct of Research published by The Finnish Advisory Board on Research Integrity (TENK; www.tenk.fi). Written informed consent was obtained from all participants in the study. The participants had the right to withdraw from the study at any stage and/or to deny the use of their samples and medical records without explanation. The project design was approved by the local Research Ethics Committee at Pirkanmaa Hospital District and by the National Supervisory Authority for Welfare and Health.

Permissions for the familial sample collection and for the use of data stored in the Finnish Cancer Registry were granted by the Ministry of Social Affairs and Health (license no. 59/08/95). Permission to use tissue samples from prostate cancer patients for medical research purposes was granted by the National Authority for Medicolegal Affairs in 2006 and extended in 2010 (license no. 5569/32/300/05). Permission to collect and use samples from unselected prostate cancer patients living in the Pirkanmaa area was granted by the Research Ethics Committee of the Pirkanmaa Hospital District in 2003 and extended in 2010 (license no. R03203). Permission to use the samples and clinical data from prostate cancer patients treated at the Hatanpää City Hospital was granted by the Institutional Review Board of the City of Tampere (license no. 8595/403/2005). Permissions to use the breast and colon cancer sample collections in Helsinki were granted by the Ethics Committees of the Departments of Obstetrics, Gynaecology and Oncology at Helsinki University Central Hospital and by the Ministry of Social Affairs and Health.

4.2 Human cell lines and xenografts (I)

Two cell lines representing normal prostate epithelium were included in the first study. PrEC Prostate Epithelial Cells were obtained from Lonza (Walkersville, MD, USA). EP156T, a primary prostate epithelial cell line immortalized by human telomerase reverse transcriptase (hTERT), was provided by Dr. Varda Rotter (Weizmann Institute of Science, Rehovot, Israel).

In addition, eight prostate cancer cell lines were analysed. DU145, PC-3, 22Rv1 and LNCaP were obtained from the American Type Culture Collection (ATCC; Manassas, VA, USA). CWR22Pc was provided by Dr. Marja Nevalainen (Thomas Jefferson University, Philadelphia, PA, USA). LAPC4 was obtained from Dr. Charles Sawyers (University of California, Los Angeles, CA, USA). VCAP and DuCaP were provided by Dr. Jack Schalken (Radboud University Nijmegen Medical Center, Nijmegen, Netherlands).

The 19 LuCaP human prostate cancer xenograft lines were obtained from Dr. Robert L. Vessella (Department of Urology, University of Washington Medical Center, Seattle, WA, USA). Xenografts, tumour tissues implanted in immunocompromised mice, represent preclinical models that have been developed to investigate the complexity of prostate cancer and to design and evaluate novel therapies.

4.3 DNA extraction (I-III)

Genomic DNA was extracted from peripheral blood lymphocytes using the Wizard® Genomic DNA Purification Kit (Promega Corporation, Madison, WI, USA). The DNA concentration was measured using an ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA).

4.4 RNA extraction (II)

Total RNA was extracted from peripheral blood samples collected in PAXgene® Blood RNA Tubes (PreAnalytiX GmbH/QIAGEN Sciences, Germantown, MD, USA) using two commercially available kits: MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit (Ambion®/Life Technologies, Carlsbad, CA, USA) and PAXgene Blood miRNA Kit (PreAnalytiX GmbH/QIAGEN). The RNA yield was quantified using an ND-1000 spectrophotometer (Nanodrop Technologies). An Agilent 2100 Bioanalyzer and the Agilent RNA 6000 Nano Kit (Agilent Technologies, Santa Clara, CA, USA) were used to assess RNA integrity and quality.

4.5 Sequencing (I, II)

4.5.1 Direct DNA sequencing (I, II)

Standard Sanger sequencing was used in Study I to confirm the *HOXB13* variant p.G84E (rs138213197) in prostate and breast cancer cases and controls. In Study II, family members from 41 HPC families were genotyped by sequencing to explore the co-occurrence of *ZNF652* variants rs116890317 and rs79670217, *EFCAB13* variant rs118004742 and *HDAC4* variant rs73000144 with disease phenotype. In addition, the coding exons of the *ANO7* gene were sequenced to screen for prostate-cancer-associated variants. The primer sequences are listed in Table 7. Additional sequencing was performed by the Department of Medical Genetics, Genome-Scale Biology Research Program of the University of Helsinki, where the whole coding region of *HOXB13* was sequenced from colorectal cancer cases and controls for p.G84E genotyping and LOH analysis.

To prepare the PCR-amplified products for sequencing, the ExoSAP-IT™ PCR Clean-Up Protocol (USB Corporation/GE Healthcare, Cleveland, OH, USA) was applied. Sequencing was performed using the ABI PRISM® BigDye™ Terminator Cycle Sequencing Ready Reaction Kit and the ABI 3130xl Genetic Analyzer (Applied Biosystems/Life Technologies, Carlsbad, CA, USA) following the manufacturer's recommendations. Different versions of Sequencher software (GeneCodes Corporation, Ann Arbor, MI, USA) were used for sequence analysis.

The identified variants were named according to the following reference sequences (RefSeq ID): *ANO7* (NM_001001891), *EFCAB13* (NM_152347), *HDAC4* (NM_006037), *HOXB13* (NM_006361) and *ZNF652* (NM_001145365). The reference sequences were obtained from NCBI's Reference Sequence Database (RefSeq; www.ncbi.nlm.nih.gov/refseq/).

Table 7. Sequences (5'→3') of the primers used in Sanger sequencing.

Gene	Exon / Target	Forward primer	Reverse primer
ANO7	1	AGCTGTGCTGGGCACCTC	CCTAGAGTCCAACGCTCCAC
	2	GTCTCACCCATCCCCTCTCT	GACCTCTCAAGTCGCACCAC
	3	GGGTGGGTGTAGTTGTCGAG	TGGCTACTGAGGAGGCTACC
	4	GGCCACTGCCACTTAGCC	ATGGGTCACTGAGTGGATGC
	5	ACGGCTACAGAAATGCCAGT	CAGCTGAACGCAGTGTGTG
	6	TTTGCAAACTTGCACAACCT	CCAAGTTCCGCTCACTCATT
	7	GAGCCAGCTGCTTCTCCTG	GATCCTCAGAGCCAGGTCAG
	8	ATGTGCATGTGCGGTGGT	TTCCCAGCAAGAGACGCTAC
	9	GCCCCTGCACCTACAACAG	TACAACCTGACACCAAGCTG
	10	CCTGGGTTCCTGATGGTG	AAGCACCAGCTGTCTGCAC
	11	CAAGGGAGAGAGAGGACAAGG	TCATCCCCGACTCTCAAATC
	12 - 13	AAATACACAGTCGGGGGATGT	GGGGAGTGAGGGTTCTGTG
	14	CCCAAGACACCGTGAAGG	AGAGGCCTAACGGGAGACAG
	15	GTCGGGCAACACCCTTCT	GGCCATGTGTGTCAGTGAGT
	16	GCAAGGTGGTCCTAGGAGAG	GCTGGATGACGCCTGGTA
	17	CTTCCTCCAGGGCAGGTG	CAAATCAAAGCTCGAATGGA
	18	GGCCAGCTTTGAGACAAGAA	CTGCTACTGCCAGGTGCTC
	19	CCTTCAATTGCAAAGCAACA	CAGCACATTTCAGGGCAGAT
	20	GTGACTGGAGAAGCTGGTTG	GCCTCACGTTGCTGATGAC
	21	GGTCATCAGCAACGTGAGG	CAAAGCTCCGTCCCTTACC
	22	AGAGAGCGAAATGGTGGAAA	CTCACCACGTGCTCGAATTA
	23 - 24	CTGCACAGCTGCTTTCTGAC	AAGTCAGACTAGGGCCAGGA
	25	GCTCCTGGCCCTAGTCTGA	AGAGATGAGGCACAACAGCA
EFCAB13	rs118004742	CACTGCTGAAGTGATTTATATTTTTGT	CAAATTGACCCTCCTTCCAG
HDAC4	rs73000144	GTCGCAAGCTAACGAGCAG	TGGGGTCATTTCAAGCTCAT
HOXB13	1	CGAGCTGGGAGCGATTTA	AGCACCAAGCTCATCCTCAC
HOXB13	rs138213197	CTGTCAACTATGCCCCCTTG	GCGGCTGGGGTACTCTTC
ZNF652	rs116890317	TCTATCCAGTAGGTCATCTTAGGG	GGGCACATGGTAGGCTATTTT
ZNF652	rs79670217	GTGAAGGTGAGGGTCAATGC	AGGGTTTGAGTTGTTACCCTAAA

4.5.2 Targeted DNA re-sequencing and variant selection (II)

The targeted re-sequencing of the prostate-cancer-associated loci 2q37 and 17q11.2-q22 was performed by paired-end next-generation sequencing at the Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki. The sequenced regions spanned approximately 6.8 Mb and 21.6 Mb, respectively. SeqCap EZ Choice array probes (Roche NimbleGen, Madison, WI, USA) were used to capture the target regions, which were then sequenced on a Genome Analyzer IIx (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. The Variant Calling Pipeline (VCP) developed in FIMM (Sulonen et al. 2011) was applied for read alignment and variant calling.

Variant selection consisted of multiple steps and aimed to identify the variants most likely involved in prostate cancer susceptibility. First, only variants that co-segregated with affection status in the analysed families were retained. Variant annotation was then performed to exclude variants that resided within intergenic regions or in non-coding genes. The final filtering steps consisted of pathogenicity predictions, followed by literature and database queries, which are described in more detail in Chapter 4.8.

4.5.3 RNA sequencing (II)

Whole-genome gene expression was inspected by massively parallel paired-end RNA sequencing. The cDNA sequencing library was prepared according to Illumina's protocol (Sample Preparation Guide for mRNA Sequencing, Part #1004898, Rev. A, Illumina). First, the Dynabeads® mRNA Purification Kit (Invitrogen/Thermo Fisher Scientific, Waltham, MA, USA) was used to enrich the poly-A-containing mRNA molecules from the total RNA sample. The enrichment was followed by the mRNA fragmentation and random hexamer-primed cDNA synthesis of the first cDNA strand using reverse transcriptase. The second cDNA strand was synthesized using DNA Polymerase I and RNaseH. After end reparation and the addition of sequencing adapters, the cDNA fragments were amplified by PCR and sequenced using the Illumina HiSeq 2000 sequencing platform (Illumina). The transcriptome quantification was performed at Beijing Genomics Institute (BGI Hong Kong Co., Tai Po, Hong Kong).

The quality of the RNA sequencing reads was checked with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Read alignment was performed with the splice junction mapper TopHat2 (Trapnell et al. 2009) using

GRCh37/hg19 as the reference genome. HTSeq software (Anders et al. 2015) was used to determine the number of reads that overlapped genes. The DESeq package for R (Anders & Huber 2010) was used to transform the raw read counts into comparable expression values via normalization. Genes with very low (normalized read counts of <20) or no expression were excluded from further analysis.

4.6 High-throughput genotyping (I-III)

4.6.1 TaqMan SNP genotyping (I)

The frequency of the *HOXB13* variant p.G84E (rs138213197) was determined in a total of 12,502 samples, including prostate and breast cancer cases and controls, prostate cell lines and xenografts, using the Custom TaqMan® SNP Genotyping Assay for rs138213197 (Applied Biosystems/Life Technologies) and the ABI Prism 7900HT Sequence Detection System (Applied Biosystems/Life Technologies). The genotyping was performed in a 384-well format with duplicate samples and four negative controls per plate. The Freedom EVO® pipetting robot (Tecan Group Ltd, Männedorf, Switzerland) was used for plate preparation. The primer and probe pairs for rs138213197 were designed with the Custom TaqMan® Assay Design Tool available on the Applied Biosystems website.

4.6.2 Sequenom MassARRAY genotyping (I, II)

To validate the 58 prostate-cancer-related candidate variants that were identified in the targeted re-sequencing step of Study II, these variants were genotyped in a total of 2,216 samples using the Sequenom MassARRAY® System and the iPLEX® Gold Assays (Sequenom, San Diego, CA, USA). The BPH samples included in Study I were genotyped for the *HOXB13* mutation p.G84E with the same method. The MassARRAY® System is based on locus-specific PCR using mass-modified dideoxynucleotide terminators. The SNP allele at the locus of interest is identified by measuring the distinct mass of the incorporated terminator with Matrix-Assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) mass spectrometry. This technology enables the simultaneous analysis of up to 40 different SNPs in a single multiplex reaction and is therefore suitable for rapid SNP validation in a large sample set. All of the genotyping reactions were performed at FIMM following the manufacturer's instructions. To ensure genotyping quality, duplicate samples, negative controls and success rate checks were included in the protocol. TyperAnalyzer Software (Sequenom) was used for genotype calling.

4.6.3 Genome-wide SNP scan (III)

Genome-wide SNP genotyping was performed for a total of 142 samples using the Infinium® HD Assay and the HumanOmniExpress-12 v1.0 BeadChip Kit (Illumina). This microarray enabled the simultaneous genotyping of more than 733,000 optimized tag SNPs with a mean spacing between markers of 4.0 kb. Sample preparation and SNP genotyping were conducted at FIMM as recommended by the manufacturer. The GenomeStudio Software (Illumina) and the PennCNV algorithm (Wang et al. 2007) were used for CNV calling. CNVs encompassing fewer than three SNP markers were omitted from further analysis.

4.6.4 TaqMan copy number variation analysis (III)

To validate the four CNVs that showed enrichment in familial prostate cancer patients, a total of 665 samples were genotyped using the predesigned TaqMan® Copy Number Assays Hs05836821_cn (2q34), Hs03480483_cn (3p11.1), Hs00434275_cn (5p13.3) and Hs03692888_cn (8p23.2) (Applied Biosystems/Life Technologies). An additional 210 samples were genotyped to study the co-occurrence of these four CNVs with prostate cancer. Quadruple reactions were prepared on 384-well plates for each of the 878 samples, and the TaqMan® RNaseP Reference Assay (Applied Biosystems/Life Technologies) was included as an internal standard. The detection of the real-time PCR products was performed on the ABI PRISM 7900HT Sequence Detection System. Copy numbers were determined with the CopyCaller™ Software (Applied Biosystems/Life Technologies).

4.7 Expression quantitative trait loci (eQTL) mapping (II)

Prostate-cancer-associated regulatory variants within the 2q37 and 17q11.2-q22 loci were localized using the eQTL mapping method. In this analysis, the whole-genome gene expression data obtained by RNA sequencing were combined with SNP genotypes derived from targeted DNA re-sequencing. Only *cis*-eQTLs within a 2 Mb window were considered, i.e., the regulatory SNP had to be located within 1 Mb upstream or downstream of its target gene. Two different eQTL mapping strategies were applied. In the first approach, only genes that were differentially expressed (DE) between prostate cancer patients and controls were included in the analysis. The second, modified approach exploited variants with a known, previously reported association with prostate cancer. These variants were identified in a recent GWAS (Eeles et al. 2013) that was performed on a large multinational prostate cancer cohort. The samples were analysed using a custom-made iSelect SNP genotyping array (Illumina), the iCOGS array, designed by the international Collaborative Oncological Gene-environment Study (iCOGS) Consortium (Sakoda et al. 2013).

To test for SNP-gene associations, a new type of generalized Mann-Whitney test (Fischer et al. 2014) was used. This directional test, which is based on probabilistic indices, is less sensitive to outliers than are traditional linear tests. This analysis was performed using the R packages gMWT and GeneticTools available on the Comprehensive R Archive Network (CRAN; https://www.cran.r-project.org). Permutation tests with 1,000 permutations were used to determine the p values for the observed eQTLs with a significance level of $p \leq 0.005$.

4.8 Bioinformatics (I-III)

To evaluate the pathogenicity and possible clinical significance of the identified variants, several *in silico* pathogenicity predictors were employed in Studies I and II and in the assessment of the *ANO7* variants. These predictors included PON-P (Olatubosun et al. 2012), PolyPhen-2 (Adzhubei et al. 2013), SIFT (Kumar et al. 2009), MutationTaster (Schwarz et al. 2014), SNAP (Bromberg & Rost 2007), PhD-SNP (Capriotti et al. 2006), PANTHER (Mi et al. 2010), SNPs&GO (Calabrese et al. 2009) and PROVEAN (Choi et al. 2012). Further *in silico* predictions in Study I assessed the biochemical properties of the HOXB13 protein. The amino acid sequence surrounding glycine 84 was examined using NetSurfP (Petersen et al. 2009) and SABLE 2 (Adamczak et al. 2005). I-Mutant-3 (Capriotti et al. 2008), MuPro

(Cheng et al. 2006) and iPTREE-STAB (Huang et al. 2007b) were used to investigate the impact of the p.G84E variant on HOXB13 protein stability.

In Studies II and III, literature and database searches were performed to obtain information on the frequencies and disease associations of the identified SNP and copy number variants. MAFs were retrieved from the dbSNP database for the 152 candidate variants that had been predicted to be pathogenic in Study II. The COSMIC (Forbes et al. 2015) and DDPC (Maqungo et al. 2011) databases were used to discover genes with a previously reported association with prostate cancer. The novelty of the CNVs identified in Study III was investigated by comparing the CNVs with previously reported CNV data available in DGV. GRCh37/Hg19 was used as the reference assembly. The NCBI RefSeq database was used to uncover genes that overlapped with the CNV loci. For intergenic CNVs, the nearest gene was located using the BEDTools suite (Quinlan & Hall 2010). To find evidence for previous association with cancer, all of the CNV-linked genes were queried for overlap against genes listed in the OMIM database.

The biological functions of the candidate genes identified in Studies II and III were examined using the Gene Ontology (GO) and pathway data. In Study II, GO data were gathered via Ensembl BioMart v.65 (Flicek et al. 2013), and pathway data were retrieved from Pathway Commons (Cerami et al. 2011), KEGG (Kanehisa & Goto 2000) and WikiPathways (Pico et al. 2008). In Study III, WebGestalt2 (Wang et al. 2013) was used to analyse the participating pathways. The false discovery rate was reduced by adjusting the *p* values using the Benjamini-Hochberg method. In addition, ENCODE information (ENCODE Project Consortium 2012) for the non-coding variants and eQTLs identified in Study II was obtained using RegulomeDB (Boyle et al. 2012).

In Study III, the accumulation of CNVs in prostate cancer families was investigated using a family-based enrichment analysis. For each family, the total number of prostate cancer patients, the number of genotyped patients and the number of patients with CNV were determined. The percentages of CNV carriers from the total number of patients and from the genotyped patients were then calculated. For a CNV to be enriched in a family, the proportion of carriers had to be ≥50%.

4.9 Statistical analysis (I-III)

Hardy-Weinberg equilibrium (HWE) tests were performed using PLINK (Purcell et al. 2007). The association between the identified variants and prostate cancer risk (and in Study I, breast and colorectal cancer risk) was examined using PLINK and GraphPad Prism 5.02 (GraphPad Software, La Jolla, CA, USA). A two-sided Fisher's exact test implemented in PLINK was used to study the statistical significance of the observed associations. PLINK was also used to calculate the odds ratios (ORs) and the corresponding 95% confidence intervals (CIs).

To examine the association between the *HOXB13* variant p.G84E and overall survival in Study I, the Cox model was employed. Log-rank and Gehan-Breslow-Wilcoxon tests were used to evaluate the statistical significance of the survival differences between carriers and non-carriers.

In Study II, linkage disequilibrium (LD) was examined among the 13 variants on chromosomes 2 and 17 that associated significantly with prostate cancer. Additional LD mapping included these 13 significant variants and the prostate-cancer-associated variants identified by the iCOGS GWAS (Eeles et al. 2013). The Haploview program (Barrett et al. 2005) was used to calculate the r^2 values between the variants and to view the haplotype blocks.

To evaluate the CNV distribution and median CNV lengths between patients and controls in Study III, the Wilcoxon test implemented in R was employed. Fisher's exact test was used to compare the frequencies of CNV carriers between patients and controls. If the Fisher's exact test failed to estimate the *p* values and ORs due to zero denominators, these statistics were calculated using the Visualizing Categorical Data (VCD) package (Meyer et al. 2014) implemented in R.

5 Summary of the Results

5.1 Novel prostate-cancer-associated sequence variants at the 2q37 and 17q11.2-q22 loci (II)

To explain the previously reported linkage between prostate cancer and the 2q37 and 17q11.2-q22 loci (Cropp et al. 2011), a detailed genomic characterization of these two regions was performed. The fine-mapping consisted of the targeted re-sequencing of the regions in 68 samples from 21 Finnish high-risk HPC families, followed by variant characterization and prioritization. In total, 107,479 unique sequence variants were detected across all samples by sequencing, and 40,612 of these variants were observed to co-segregate with prostate cancer in the analysed families. Following annotation, 24,813 variants located within protein-coding genes were retained. Pathogenicity predictors classified only 152 variants as possibly pathogenic. Additional information on these variants was obtained by literature review and extensive database searches, and priority was given to rare variants (MAF <0.05) located in known or suspected prostate-cancer-associated genes or genes functionally similar to them. Finally, 58 variants in 35 candidate genes were selected for validation in 2,216 study subjects using the Sequenom MassARRAY genotyping system. In a subsequent case-control association analysis, variant allele frequencies were compared between familial index cases or unselected prostate cancer patients and population controls to identify novel variants that may predispose their carriers to prostate cancer.

Table 8 lists variants that associated with prostate cancer with a nominal *p* value of 0.05 in either the familial or the unselected sample set and had an OR >1.0. These included six variants in four different genes: *ZNF652, HOXB3 (Homeobox B3)* and *EFCAB13 (EF-Hand Calcium Binding Domain 13)* on chromosome 17 and *HDAC4* on chromosome 2. The strongest association with an increased prostate cancer risk in both the familial and unselected sample sets was observed for two intronic *ZNF652* variants, rs116890317 and rs79670217. The OR for rs116890317 was 7.8 among the familial samples (95% CI 3.0-20.3, p = 3.3 x 10^{-5}) and 3.3 among the unselected samples (95% CI 1.4-7.5, p = 0.003). The risk effect of rs79670217 was less pronounced, as reflected by the ORs of 1.9 for the familial set (95% CI 1.2-3.1, *p* = 0.009) and 1.6 for the unselected set (95% CI 1.2-2.2, *p* = 0.002). The two non-coding *HOXB3* variants, rs10554930 and rs35384813, were common among all patients and controls, with a variant allele frequency of >20%. They associated with a modest prostate cancer risk among the familial patient group only (OR = 1.4, 95% CI 1.1-1.8, *p* = 0.010-0.013). The *HDAC4* missense variant rs73000144 (c.958C>T, p.Val320Ile) was observed at a very low frequency. Only three familial patients (1.6%), seven unselected patients (0.6%) and one control individual (0.1%) were identified as heterozygous carriers. Rs73000144 associated with prostate cancer risk within the familial sample set (OR = 14.6, 95% CI 1.5-140.2, *p* = 0.018) but not within the unselected sample set (OR = 5.9, 95% CI 0.7-47.9, *p* = 0.078). Borderline association with increased prostate cancer risk was also obtained for the *EFCAB13* nonsense variant rs118004742 (c.1638T>G, p.Tyr546Ter) but among the familial patients only (OR = 1.8, 95% CI 1.0-3.1, *p* = 0.048). Among the unselected patients, no risk effect for rs118004742 was observed (OR = 1.1, 95% CI 0.8-1.6, *p* = 0.637).

Table 8. Variants associated with prostate cancer (*p* < 0.05) at 2q37 or 17q11.2-q22.

SNP Id	Variant allele frequency					
Gene (Locus)	Controls[a] %	Patients	%	OR	95% CI	*p*
rs116890317	0.39	Familial[b]	2.96	7.8	3.0 – 20.3	3.3×10^{-5}
ZNF652 (17q21.3)		Unselected[c]	1.27	3.3	1.4 – 7.5	0.003
rs79670217	3.56	Familial	6.65	1.9	1.2 – 3.1	0.009
ZNF652 (17q21.3)		Unselected	5.66	1.6	1.2 – 2.2	0.002
rs10554930	21.3	Familial	27.5	1.4	1.1 – 1.8	0.010
HOXB3 (17q21.3)		Unselected	24.1	1.2	1.0 – 1.4	0.034
rs35384813	20.8	Familial	26.7	1.4	1.1 – 1.8	0.013
HOXB3 (17q21.3)		Unselected	23.2	1.1	1.0 – 1.3	0.073
rs73000144	0.06	Familial	0.80	14.6	1.5 – 140.2	0.018
HDAC4 (2q37.2)		Unselected	0.33	5.9	0.7 – 47.9	0.078
rs118004742	2.73	Familial	4.79	1.8	1.0 – 3.1	0.048
EFCAB13 (17q21.3)		Unselected	3.00	1.1	0.8 – 1.6	0.637

[a] Male population controls (n = 914)

[b] Familial index patients (n = 186)

[c] Unselected prostate cancer patients (n = 1096)

According to MutationTaster, one of the *in silico* pathogenicity predictors used, both of the *ZNF652* variants and the *HDAC4* variant rs73000144 were defined as benign polymorphisms. Rs73000144 was also classified as benign or neutral by two additional predictors, PolyPhen-2 and PON-P. In contrast, the two *HOXB3* variants and the *EFCAB13* variant rs118004742 were predicted to be pathogenic by MutationTaster.

5.2 eQTL analysis of the 2q37 and 17q11.2-q22 loci (II)

The 2q37 and 17q11.2-q22 loci were also mapped for *cis*-acting regulatory variants that may control the expression of prostate-cancer-associated genes. To identify such variants and their target genes, whole-transcriptome sequencing was performed, followed by eQTL analysis restricted within these two regions of interest.

The first, traditional eQTL mapping strategy exploited differential gene expression (DE) profiles between prostate cancer patients and unaffected individuals. The DE analysis was performed on 173 genes at 2q37 and 761 genes at 17q11.2-q22. Significant differences in expression levels ($p < 0.05$) between patients and controls were observed for eight genes: three genes on chromosome 2 and five genes on chromosome 17. In the targeted *cis*-eQTL analysis, a total of 54,919 SNPs were tested for association with these eight DE genes within a 2 Mb detection window. To minimize the number of false-positive results, the significance level for SNP-gene associations was set to $p \leq 0.005$. Altogether, 272 candidate regulatory SNPs were identified for six DE genes, with three genes on each chromosome. The majority (87%) of the regulatory SNPs (237 out of 272) were located at 2q37, whereas only 35 SNPs (13%) were found at 17q11.2-q22. To evaluate the regulatory potential of the identified SNPs, ENCODE data (ENCODE Project Consortium 2012) was incorporated, and the strongest evidence of functionality was obtained for two SNPs: rs12620966 on chromosome 2 and rs11650354 on chromosome 17 (Table 9). Rs12620966 targets *AGAP1 (ArfGAP With GTPase Domain, Ankyrin Repeat And PH Domain 1)*, whereas rs11650354 is associated with differential expression levels of *TBKBP1 (TANK-Binding Kinase 1-Binding Protein 1).*

The modified eQTL analysis applied a set of pre-filtered SNPs that had been associated with prostate cancer in a previous study (Eeles et al. 2013). This set consisted of 12 SNPs at 2q37 and 22 SNPs at 17q11.2-q22. The effect of these SNPs was investigated on the expression of 144 genes on chromosome 2 and 160 genes on chromosome 17. Only one prostate-cancer-associated ($p \leq 0.005$) *cis*-eQTL was identified on chromosome 2, whereas on chromosome 17, a total of 36 candidate eQTLs were found. ENCODE data (ENCODE Project Consortium 2012) suggested the strongest regulatory potential for four SNPs at 17q11.2-q22, listed in Table 9. Information on the chromosome 17 variant rs4793976 is also included, although no ENCODE data were available for this SNP. Rs4793976 targets *SPOP (Speckle-Type POZ Protein)*, a known prostate cancer candidate gene frequently mutated in a subclass of prostate cancers (Barbieri et al. 2012).

Table 9. Summary of *cis*-eQTLs at 2q37 and 17q11.2-q22 with the strongest evidence of regulatory potential. The first two eQTLs were identified by traditional analysis, and the last five by a modified approach.

Variant	Chr	Target gene[a]	Distance (kb)[b]	*p*	Regulome score[c]	Evidence for TF binding[d]	Evidence for Open chromatin[e]
rs12620966	2	*AGAP1*	634.6	0.002	2a	ChIP-seq, DF, PWM	DNase-seq
rs11650354	17	*TBKBP1*	32.6	0.004	1f	ChIP-seq	DNase-seq
rs4796751	17	*DHX58*	125.9	0.001	1f	-	DNase-seq, FAIRE
		MLX	591.5	0.004	1f	-	DNase-seq, FAIRE
rs4796616	17	*JUP*	62.0	0	1f	ChIP-seq	DNase-seq
rs4793943	17	*ZNF652*	699.6	0.003	2b	ChIP-seq	DNase-seq
rs16941107	17	*ARL17B*	460.6	0.004	2b	ChIP-seq	DNase-seq
rs4793976	17	*SPOP*	895.7	0.002	-	-	-

[a] *DHX58 = DEXH Box Polypeptide 58, MLX = MLX, MAX Dimerization Protein, JUP = Junction Plakoglobin, ARL17B = ADP-Ribosylation Factor-Like 17B*

[b] Distance of SNP from target gene

[c] 1f = likely to affect binding and linked to expression of a target gene, 2a/2b = likely to affect binding

[d] ChIP-seq = Chromatin Immunoprecipitation sequencing, DF = Digital DNase I Footprinting, PWM = Position Weight Matrix matching

[e] DNase-seq = Deoxyribonuclease I (DNase I) hypersensitive sites sequencing, FAIRE = Formaldehyde-Assisted Isolation of Regulatory Elements

5.3 The *HOXB13* variant p.G84E is associated with increased prostate cancer risk (I)

The p.G84E variant in the novel prostate cancer candidate gene *HOXB13* at 17q21.3 has been reported to be significantly associated with an increased risk of hereditary prostate cancer (Ewing et al. 2012). To determine the frequency of the p.G84E variant among Finnish familial and unselected cancer cohorts and to investigate the possible association of the variant with cancer in Finland, we genotyped a total of 13,919 samples representing prostate, breast and colorectal cancer patients, patients with BPH and unaffected controls. In addition, the association of the p.G84E variant with selected clinical characteristics of prostate cancer was studied. The median survival time after prostate cancer diagnosis was also compared between carriers and non-carriers.

The frequencies of the p.G84E carriers among prostate cancer patients (n = 4,571) and unaffected male controls (n = 5,467) are summarized in Table 10. The highest carrier frequency (8.4%) was detected among index patients of high-risk HPC families. The variant was less common among unselected prostate cancer patients (3.6%), followed by the ERSPC screening trial patients (2.2%). The lowest carrier frequencies were obtained for the two control groups: 1.0% for population controls and 0.3% for the ERSPC controls. Case-control association analyses demonstrated that the p.G84E variant contributed significantly to increased prostate cancer risk ($p < 0.05$) among all patient groups. In particular, the risk of familial prostate cancer was elevated (OR = 8.8, 95% CI 4.9-15.7, $p = 2.3 \times 10^{-18}$), but the association was statistically significant among the unselected (OR = 3.3, 95% CI 2.2-5.7, $p = 1.8 \times 10^{-8}$) and the ERSPC screening trial patients (OR = 2.1, 95% CI 1.2-3.6, $p = 0.0046$) as well. Even stronger associations were observed when prostate cancer patients were compared to the ERSPC controls (lower part of Table 10).

A case-case association analysis revealed a connection between the p.G84E variant and earlier age at diagnosis. Variant carriers were more likely than were non-carriers to develop prostate cancer before the age of 55 years (OR = 2.0, 95% CI 1.3-3.0, $p = 0.0008$). In addition, the p.G84E variant was found to correlate with a higher serum PSA concentration (≥20 ng/ml) at the time of diagnosis (OR = 1.4, 95% CI 1.1-1.9, $p = 0.006$). Instead, statistical evidence for an association with higher tumour grade (Gleason score ≥8 vs. ≤6) or prostate cancer progression as indicated by elevated PSA (present vs. absent) was not obtained. The p.G84E variant did not correlate with decreased survival time either (HR = 1.16, 95% CI 0.9-1.5). An analysis of the BPH cohort revealed that carriers of the p.G84E variant had an increased risk of developing prostate cancer compared to that of non-carriers (OR = 4.6, 95% CI 1.3-16.2, p = 0.011).

In the breast cancer cohort, 1.9% of the familial patients, 1.5% of the unselected patients and 1.1% of the population controls carried the p.G84E variant. Similar carrier frequencies were obtained for the colorectal cancer cohort, in which 1.6% of the unselected colorectal cancer patients and 0.9% of the population controls were identified as carriers. Case-control association analyses revealed that differences in carrier frequencies were not significant for either of these cancer cohorts. In addition, the LOH analysis of the p.G84E-positive colorectal tumours produced normal results, with no indication of allelic imbalance.

Table 10. Association of the p.G84E variant with prostate cancer risk. A significant association with increased cancer risk (p < 0.05) was observed in all comparisons between prostate cancer patients and the two control groups (population controls and ERSPC* controls).

Sample set	Carrier frequency	OR	95% CI	*p*
Male population controls	9/923 (1.0%)	1.0		
Familial index patients	16/190 (8.4%)	8.8	4.9 – 15.7	2.3×10^{-18}
Unselected prostate cancer patients	114/3197 (3.6%)	3.6	2.2 – 5.7	1.8×10^{-8}
Prostate cancer patients from ERSPC	26/1184 (2.2%)	2.1	1.2 – 3.6	0.0046
ERSPC controls	13/4544 (0.3%)	1.0		
Familial index patients	16/190 (8.4%)	33.1	19.4 – 56.5	1.8×10^{-89}
Unselected prostate cancer patients	114/3197 (3.6%)	13.4	8.9 – 20.3	6.2×10^{-57}
Prostate cancer patients from ERSPC	26/1184 (2.2%)	8.0	4.9 – 12.9	1.1×10^{-23}

* ERSPC = the European Randomized Study of Screening for Prostate Cancer

To evaluate the pathogenicity of the *HOXB13* variant p.G84E, six different *in silico* tolerance predictors, SIFT, PolyPhen-2, PON-P, PHD-SNP, SNAP and Panther, were employed. With the exception of PON-P, the programs predicted the variant to be pathogenic. NetSurfP and SABLE2 were used to investigate the amino acid sequence environment flanking the variant. The results suggest that glycine 84 was located in a hydrophobic region buried inside the HOXB13 protein. The effect of p.G84E on protein stability remained unresolved, as the three stability predictors that were applied, I-Mutant-3, MuPro and iPTREE-STAB, gave contradictory results.

5.4 *ANO7* may contribute to familial prostate cancer risk

The prostate cancer candidate gene *ANO7* is located at 2q37.3, near the subtelomeric region of the long arm of chromosome 2. Due to its distal position, this locus was not covered by the targeted re-sequencing described in Chapter 5.1. Therefore, to identify prostate-cancer-associated variants in the *ANO7* gene, we sequenced all 25 coding exons of the gene together with the flanking intronic splice sites. Altogether, 37 of the most representative Finnish high-risk HPC families were selected for mutation screening. The number of analysed prostate cancer patients ranged from 78 to 105 and a median of three patients per family were genotyped.

In total, 23 *ANO7* sequence variants were detected in the screened samples. Only one variant, a 38-bp deletion in intron 6 (c.717-69del38), was novel. This variant was detected in all prostate cancer patients (n = 6) from two families. The genotyping of an additional 20 unaffected family members revealed that the deletion did not co-segregate with disease phenotype because unaffected deletion carriers (n = 4) were identified in both families. Fifteen of the variants were exonic, including eight missense variants, five silent variants, one nonsense variant and one frameshift variant. *In silico* pathogenicity analyses using five different predictors (SIFT, PROVEAN, MutationTaster, SNPs&GO and PolyPhen-2) were performed for eight previously reported variants whose MAF was <1%. Three variants were predicted to be disease causing or damaging, three were classified as polymorphisms, and for two variants, conflicting results were obtained. Selected features for these eight variants are presented in Table 11.

Rs148609049, which introduced a premature stop codon in the first exon of the *ANO7* gene, was selected for further analysis. Mutations of this type likely result in the complete absence of the protein product. Moreover, rs148609049 was predicted to be disease causing by MutationTaster, and the frequency of the variant T allele was low (MAF = 0.0016). Therefore, the co-occurrence of rs148609049 was assayed in the single family in which it was detected. Altogether, 19 members of the family were genotyped, including three affected men and 16 unaffected relatives. The analysis revealed that the variant did not co-segregate with the disease. Nine unaffected individuals were identified as heterozygous carriers, and the variant was observed in a homozygous state in one unaffected male.

Table 11. A summary of eight *ANO7* candidate variants with MAF <1% identified in Finnish HPC patients.

SNP Id	Variation	Function	MAF*	Prediction	No. of carriers (No. of families)
rs148609049	c.88C>T	p.Arg30Ter	T = 0.0016	Disease causing	2/105 (1/37)
rs34069570	c.208G>A	p.Asp70Asn	A = 0.0082	Polymorphism	6/78 (5/30)
rs77559646	c.471+5G>A	intronic	A = 0.0068	Damaging / Benign	12/78 (5/30)
rs77482050	c.676G>A	p.Glu226Lys	A = 0.0050	Polymorphism	7/93 (3/33)
rs761832893	c.1042G>A	p.Ala348Thr	NA	Damaging	1/79 (1/30)
rs757940063	c.1051+14G>A	intronic	NA	Polymorphism	1/79 (1/30)
rs747084134	c.1121_1122insG	p.Val3755Glyfs	NA	Disease causing	1/79 (1/30)
rs181722382	c.2792T>C	p.Leu931Pro	C = 0.0008	Benign / Damaging	1/95 (1/33)

* MAF source: 1000 Genomes

5.5 Germline copy number variants and familial prostate cancer risk (III)

Highly penetrant, rare SNPs and common, low-risk SNPs explain less than half of the inherited prostate cancer risk. To survey the missing heritability, we searched for germline copy number variants (CNVs) in 142 members of 31 Finnish high-risk HPC families using a genome-wide SNP array containing more than 733,000 markers. The findings were further validated by real-time quantitative PCR (qPCR) in a larger sample set consisting of 189 familial index patients and 476 population-matched, unaffected male controls.

The PennCNV algorithm used in CNV calling identified a total of 2,575 CNVs, approximately 18 variants per individual sample. Nearly all variants (94.6%) were heterozygous, and 46 of them (1.78%) were novel. The CNVs were located at 544 different genomic loci distributed along the 22 autosomal chromosomes. Deletions represented 72% of the variants and were thus more than twice as frequent as duplications (28%). In general, duplications were larger than deletions. While the median length of deletions was <10 kb, duplications spanned >20 kb. However, the differences in CNV distribution and median CNV length between prostate cancer patients and unaffected relatives were not significant.

Through a family-based enrichment analysis, 63 CNVs were identified that were over-represented in patients from 26 HPC families. These variants were further prioritized to specify the most relevant CNVs for validation. In the prioritization process, genes with a known association with prostate cancer or genes with a potential biological role in cancer-related pathways were favoured. In addition, CNVs that were detected in more than one family and clustered predominantly in patients were preferred.

The four qPCR-validated CNVs included intronic deletions overlapping the *ERBB4 (V-Erb-B2 Avian Erythroblastic Leukemia Viral Oncogene Homolog 4), EPHA3 (EPH Receptor A3)* and *CSMD1 (CUB And Sushi Multiple Domains 1)* genes, and an exonic duplication affecting the *PDZD2 (PDZ Domain Containing 2)* gene. Table 12 summarizes the genotyping data and the case-control association test results. The 14.7 kb deletion overlapping the *EPHA3* gene at 3p11.1 was the only CNV that associated with prostate cancer, with a nominal p value of <0.05 (OR = 2.06, 95% CI 1.18-3.61, p = 0.018). The carrier frequency for this deletion was twice as high among prostate cancer patients (11.6%) than that among unaffected controls (6.1%). GO analysis results indicated that EPHA3 is likely involved in receptor and signal transduction activities and may play a role in cell adhesion and cell-cell interactions.

Table 12. A summary of the carrier and allele frequencies for the four validated CNVs and results of the association test between the CNVs and prostate cancer risk.

CNV type (Locus)	CNV size in kb	Health status	Carrier freq. (%)	Allele freq.	OR	95% CI	*p*
EPHA3 del	14.7	unaffected	29/476 (6.1)	0.030			
(3p11.1)		affected	22/189 (11.6)	0.061	2.06	1.18 – 3.61	0.018
PDZD2 dup*	52.1	unaffected	14/476 (2.9)	0.025			
(5p13.3)		affected	11/189 (5.8)	0.045	1.82	0.97 – 3.43	0.077
ERBB4 del	25.6 – 55.7	unaffected	14/476 (2.9)	0.015			
(2q34)		affected	4/189 (2.1)	0.011	0.72	0.23 – 2.19	0.793
CSMD1 del	2.7	unaffected	49/476 (10.3)	0.051			
(8p23.2)		affected	19/189 (10.1)	0.053	1.03	0.60 – 1.76	0.892

* Not in HWE.

The *PDZD2* duplication was also more common among prostate cancer patients (5.8%) than among controls (2.9%). However, of the 14 unaffected duplication carriers, only four (0.8%) were heterozygous, while as many as ten (2.1%) were homozygous for this CNV. Among the familial index patients, the corresponding frequencies were 2.6% for heterozygotes and 3.2% for homozygotes. Due to the excess of homozygotes, the *PDZD2* duplication deviated from the Hardy-Weinberg equilibrium (HWE). The *ERBB4* and *CSMD1* deletions were not associated with increased prostate cancer risk. The *ERBB4* deletion had a higher frequency among controls (2.9%) than among patients (2.1%), and the *CSMD1* deletion was observed in approximately 10% of the analysed samples, irrespective of the affection status.

5.6 Co-occurrence of variants in prostate cancer families (I-III)

In Study I, the co-occurrence of the *HOXB13* p.G84E variant with prostate cancer was investigated in 32 families whose index cases had been identified as carriers. Altogether, the DNA samples of 141 family members were genotyped, representing 86 patients and 55 unaffected relatives. The p.G84E variant was detected in 79% (68/86) of the patients but in only 40% (22/55) of the unaffected family members, further supporting the association of the variant with increased prostate cancer risk. Complete co-segregation was established in only five families, where all of the prostate cancer patients but none of the unaffected male relatives were identified as p.G84E carriers. In the majority of families, co-segregation was incomplete, as indicated by unaffected male p.G84E carriers of advanced age and by prostate cancer patients who did not harbour the variant.

In Study II, four candidate variants located within the 2q37 and 17q11.2-q22 loci were selected for analyses of co-occurrence with prostate cancer in multiple affected families. Variant selection was based on either a strong association with prostate cancer, a high risk effect as illustrated by the OR value or predicted pathogenicity. The top four variants included rs116890317 and rs79670217 in *ZNF652*, rs73000144 in *HDAC4* and rs118004742 in *EFCAB13* (Table 8), and their co-occurrence with prostate cancer was evaluated in 41 Finnish high-risk HPC families. DNA samples obtained from 447 family members were genotyped, including 243 prostate cancer patients, 112 unaffected male relatives and 92 females. The results of the co-segregation analysis are summarized in Table 13. The two *ZNF652* variants and the *EFCAB13* variant had a higher frequency among prostate cancer patients than among unaffected family members, but the *HDAC4* variant was observed in approximately one-third of both affected and unaffected individuals. Complete co-segregation with disease status was detected in only one family for rs116890317, in one family for rs79670217 and in three families for rs118004742. Interestingly, in ten families that were positive for rs116890317, 12 out of 21 prostate cancer patients (57%) also carried the *HOXB13* variant p.G84E.

Table 13. Summary of the co-segregation results for the top four candidate variants at 2q37 (*HDAC4*) and 17q11.2-q22 (*ZNF652* and *EFCAB13*).

SNP Id	Gene	No. of	Carrier frequency (%)		
		families	Patients	Unaffected males	Females
rs116890317	*ZNF652*	11/41	21/31 (67.7)	11/27 (40.7)	10/26 (38.5)
rs79670217	*ZNF652*	25/41	42/73 (57.5)	24/70 (34.3)	27/55 (49.1)
rs118004742	*EFCAB13*	19/41	31/55 (56.4)	18/44 (40.9)	18/34 (52.9)
rs73000144	*HDAC4*	3/41	3/9 (33.3)	5/12 (41.6)	2/10 (20.0)

The 14.7 kb *EPHA3* deletion was observed in index patients from 21 HPC families in Study III. To examine the co-occurrence of the deletion with affection status, the genotypes of an additional 210 family members were determined. Altogether, 56% (37/66) of the prostate cancer patients but only 36% (52/144) of the unaffected male relatives were identified as deletion carriers, suggesting a connection between the deletion and the disease. However, in each individual family, the co-segregation of the deletion with affection status was incomplete. When the cause of death among the 66 familial prostate cancer patients was examined, an interesting finding emerged. During the follow-up time of 17 to 22 years, ten patients had died of prostate cancer, and nine of them carried the *EPHA3* deletion. Whether the deletion is indeed associated with increased prostate cancer specific mortality is an interesting topic for future studies.

6 Discussion

6.1 Challenges of diagnosing clinically significant prostate cancer

Predicting prostate cancer outcomes has proven extremely challenging due to the genetic heterogeneity and phenotypic complexity of the disease. Rather than being a single clinical entity, prostate cancers represent a heterogeneous group of diseases. The majority of prostate cancers are clinically indolent, histological cancers, which grow slowly and are not life-threatening. However, in a proportion of men, the disease advances into metastatic, hormone-refractory prostate cancer, which is invariably lethal. At present, the availability of prognostic tools capable of distinguishing indolent from aggressive cancers is limited (Demichelis & Stanford 2015). Additional clinical and biochemical markers are required to improve the classification of prostate cancers into more distinct phenotypic subclasses.

The routine diagnostics of prostate cancer is currently based on the measurement of the serum PSA level, which has also been demonstrated to predict long-term cancer risk (Lilja et al. 2007). However, elevated PSA levels do not always signify cancer, and patients may have aggressive disease with a low PSA value (Alvarez-Cubero et al. 2013). A recent screening trial revealed that PSA level does not associate with a higher Gleason score and cannot, therefore, be used to reliably estimate disease aggressiveness (Boniol et al. 2015). The multifocal nature of the disease further complicates prognosis predictions. More than 70% of prostate cancers contain multiple disease foci (Villers et al. 1992, Wolters et al. 2012), which increase the risk of undergrading the cancer due to sampling bias in prostate biopsy (Fraser et al. 2015). The need for better diagnostic tools is therefore obvious.

Current efforts towards a better understanding of the biological mechanisms responsible for prostate cancer initiation, progression and metastasis have focused on the discovery of genetic biomarkers. After all, prostate cancer is known to be one of the most heritable cancers in men. A major risk factor for the disease is positive family history, and twin studies have shown that the familial risk is due to genetic, rather than environmental, components (Lichtenstein et al. 2000, Hjelmborg et al.

2014). Therefore, genetic biomarkers would most likely prove to be useful in different stages of prostate cancer: susceptibility markers could be exploited in risk assessment, diagnostic markers in early detection, prognostic markers in the estimation of disease severity, and predictive markers in the evaluation of treatment efficiency and cancer recurrence (Witte 2009, Barbieri et al. 2013).

However, the identification of appropriate genetic biomarkers has not been straightforward. The search for predisposing germline variants has been hampered by the high phenotypic variability among prostate cancer patients (Demichelis & Stanford 2015). The same disease phenotype may result from the contribution of different genetic variants in different combinations (locus heterogeneity), and familial and unselected patients may present with similar clinical characteristics (phenocopies). The reduced penetrance of some of the causative variants has led to inconsistent and confusing results in familial co-segregation studies, and the unknown effect of epidemiologic factors further complicates the deduction of true genotype-phenotype correlations (Alvarez-Cubero et al. 2013). Difficulties in the discovery of somatic variants are mainly due to intra- and interprostatic heterogeneity, multifocality and multiclonality (Fraser et al. 2015). In addition, different subclasses of prostate cancer likely result from combinations of different germline and somatic mutations in the tumour tissue, making the separation of genuinely causative driver mutations from the more insignificant passenger mutations a challenging task (Barbieri et al. 2013, Fraser et al. 2015). In multifactorial polygenic diseases, the effect size of a single variant is typically small. Consequently, large population cohorts consisting of thousands and even tens of thousands of individuals need to be analysed in order to reliably evaluate the statistical and clinical significance of rare, low-risk variants.

6.2 Contribution of known candidate genes and sequence variants to prostate cancer susceptibility in Finland (I, II)

6.2.1 Locus 17q11.2-q22

The novel prostate cancer risk gene *HOXB13*, and especially the p.G84E variant located in the first exon of the gene, play a remarkable role in prostate cancer susceptibility among North European and American populations (Smith et al. 2014). In Study I, the importance of this variant was confirmed also in Finland. As many as

8.4% of the Finnish familial prostate cancer patients and 3.6% of the unselected prostate cancer patients were shown to carry the p.G84E variant. These are the highest carrier frequencies reported to date for an individual genetic change associated with increased prostate cancer risk in Finland. Subsequent haplotype analyses have demonstrated that the *HOXB13* variant p.G84E is most likely a founder mutation (Chen et al. 2013a, Karlsson et al. 2014). Significant differences in the frequencies of the p.G84E variant across diverse populations and geographic regions have been observed. In Nordic countries, especially Finland and Sweden, the mutation is more common than in Northern America and Australia (Lynch & Shaw 2013). Interestingly, p.G84E is completely absent in China, where another *HOXB13* founder mutation, c.404G>A (p.G135E), predominates (Lin et al. 2013).

The genetic and functional mechanisms by which *HOXB13* contributes to prostate cancer susceptibility are still largely unknown. Histologically, tumours of p.G84E carriers have been reported to show features typical of benign prostatic hyperplasia (Smith et al. 2014). In the same study, *ERG* gene fusions were observed in only 22% of the tumour foci, whereas generally, *ERG* fusions are observed in 50% of prostate cancers. Therefore, it was suggested that novel molecular pathways are responsible for prostate carcinogenesis in p.G84E carriers (Smith et al. 2014). It is also possible that unforeseen functional associations explain a proportion of HOXB13-driven prostate cancers, as exemplified by *RFX6* up-regulation due to interaction between the rs339331 variant and HOXB13 (Huang et al. 2014). Such associations may be identified by a genome-wide analysis of HOXB13 binding sites. The clinical significance of the p.G84E variant in familial prostate cancer has been indisputably validated for men of European descent (e.g., Breyer et al. 2012, Xu et al. 2013). Clinical genetic testing for this variant is not yet available in Finland, but a few laboratories in Europe and the USA offer commercial genetic tests for *HOXB13* mutations (Table 1). At the moment, the usefulness of testing for the p.G84E variant remains controversial. Although the variant has been shown to associate with earlier age of onset, assessment of disease risk at an individual level is extremely difficult. Additionally, the variant does not provide information on prostate cancer prognosis nor does it affect the selection of treatment options.

ZNF652 has been associated with prostate cancer in a few previous studies. Haiman and colleagues reported the intronic *ZNF652* variant rs7210100 as associated with prostate cancer in men of African ancestry (Haiman et al. 2011). An independent European-specific risk variant, rs11650494, was described two years later (Eeles et al. 2013). This variant is not located in the *ZNF652* gene but within a long noncoding RNA (lincRNA) sequence approximately 21 kb downstream of the

ZNF652 locus. LincRNAs have been suggested to participate in the regulation of gene expression, and many of them have been associated with cancer (Kung et al. 2013). In this thesis study, neither rs7210100 nor rs11650494 was observed to correlate with increased prostate cancer risk. Instead, two novel risk variants, rs116890317 and rs79670217, were identified. Rs116890317 was rare, detected in fewer than 0.4% of controls. It associated significantly with high prostate cancer risk, and the risk effect was emphasized among the familial patients (OR = 7.8). Rs79670217 was more common and correlated with moderately increased risk (OR = 1.6-1.9), but again, the risk was higher among the familial patients. The variants were not in LD, suggesting that they contribute to prostate cancer risk independently.

Both rs116890317 and rs79670217 are located in intron 1 of the *ZNF652* gene, and the distance between them is 16.7 kb. The African-specific variant rs7210100 lies within the same intron (Haiman et al. 2011), separated from rs79670217 by 21.7 kb. This indicates that intron 1, which is >44 kb in size, may contain regulatory elements responsible for the variable expression of the *ZNF652* gene. Such elements could include intronic splicing enhancers or silencers, for example. ZNF652 is a well-characterized transcriptional repressor with multiple targets. The aberrant regulation of ZNF652 likely has widespread effects on the function of several of its target genes, which could then drive the cell towards tumourigenesis. Alternatively, rs116890317, rs79670217 and rs7210100 may be eQTLs regulating the expression of other, more distantly located genes.

6.2.2 Locus 2q37

The transcription factor encoded by *HDAC4* has been shown to repress the expression of both AR (Yang et al. 2011) and HOXB13 (Ren et al. 2009), two important TFs involved in prostate carcinogenesis. In Study II, an exonic *HDAC4* missense variant, rs73000144 (c.958C>T), was identified that was associated with increased risk of familial prostate cancer, with a nominal *p* value of 0.018. In addition, a suggestive risk effect was observed among the unselected patients. The high ORs obtained, 14.6 for familial patients and 5.9 for unselected patients, can at least partially be explained by the rarity of the variant. According to dbSNP, the MAF for the T allele is 0.0022. In our dataset, variant allele frequency was 0.06% among controls, 0.33% among unselected patients and 0.80% among familial patients.

At the protein level, the base change results in the conservative replacement of valine with another hydrophobic amino acid, isoleucine, at position 320 (p.Val320Ile). Pathogenicity predictors classified the variant as a benign polymorphism. Co-segregation analysis did not provide evidence for pathogenicity either, as the variant was equally common among affected and unaffected family members. However, only three families were included in co-segregation studies, and with small sample sizes, chance has a greater effect on the results. Perhaps rs73000144 represents a private mutation that associates only with a certain clinical subgroup of prostate cancers and is limited to specific ethnic populations. The alternative explanation, the variant being simply a passenger mutation with no effect on phenotype, cannot be excluded either. Because of the small effect size of rs73000144, the assessment of the potential role of this variant in prostate cancer predisposition requires the screening of large patient and control cohorts, preferably from several different populations.

Another candidate gene that deserves further attention is *ANO7*, which encodes a membrane protein that is putatively involved in ion and/or lipid transportation (Picollo et al. 2015). Mutations in the *ANO7* gene have recently been linked to the development and prognosis of breast cancer (Li et al. 2015). In this thesis study, the sequencing of the coding region of *ANO7* gene resulted in the identification of 23 sequence variants among Finnish prostate cancer patients. Following variant prioritization based on MAF data and pathogenicity predictions, eight candidate variants were retained. Because nonsense mutations are known to be deleterious for protein stability, structure and function, rs148609049, which introduces a stop codon in exon 1, was considered as a primary candidate. However, the variant was detected in only one family in which co-segregation with affection status could not be demonstrated. Another candidate variant that would warrant further study was rs747084134, an insertion of a guanine residue in exon 10, leading to a frameshift. Frameshift mutations are also likely to result in decreased levels of protein product, and interestingly, reduced ANO7 expression has been found to correlate with increased levels of malignancy in prostate tissue (Mohsenzadegan et al. 2013).

To investigate the potential connection of rs148609049 and rs747084134 to familial prostate cancer risk, the further characterization of these variants is required. More familial patients and their unaffected relatives need to be genotyped, followed by studies evaluating the co-segregation of the variants with affection status in relevant families. To confirm disease association, population controls need to be analysed and used as a reference group in comparison with the familial index patients. It would also be interesting to explore whether the *ANO7* variants associate with sporadic prostate cancer. This will require the genotyping of unselected prostate cancer cases. Some of these validation studies have already been undertaken (Kaikkonen E, *personal communication*).

6.3 Novel putative prostate cancer candidate genes and risk variants (II, III)

6.3.1 *EPHA3*

In Study III, genome-wide copy number variation analysis was performed to identify CNVs affecting familial prostate cancer risk. The strongest association with HPC (p = 0.018) was observed for only one CNV, a 14.7 kb deletion in intron 5 of the *EPHA3 (EPH Receptor A3)* gene at 3p11.1. Although the co-occurrence of this CNV with affection status was incomplete in the analysed families, the deletion was found to be more common among patients than among unaffected relatives. Furthermore, a suggestive correlation between prostate-cancer-specific mortality and *EPHA3* deletion was observed.

EPHA3 is a protein-tyrosine kinase belonging to the class A ephrin receptor subfamily (Stelzer et al. 2011). Ephrin receptors constitute the largest receptor tyrosine kinase family in humans, with 14 members divided into A and B classes based on their sequence homology and ligand affinity (Fox et al. 2006). Ephrin receptors are important signal transduction molecules, and their altered expression has been reported in several cancers. They have been suggested to function as either tumour suppressors or as oncogenes (Lisabeth et al. 2012). EPHA3 is involved in bidirectional signalling into neighbouring cells, and it regulates cell-cell adhesion, cytoskeletal organization and cell migration (Stelzer et al. 2011). In lung cancer, *EPHA3* is the most frequently mutated ephrin receptor gene, and somatic mutations have been detected in several other cancers as well, including colorectal cancer, melanoma, glioblastoma, hepatocellular carcinoma, pancreatic cancer and ovarian cancer (Lisabeth et al. 2012). *EPHA3* mutations contributing to prostate cancer have not been reported. Instead, another ephrin receptor gene, *EPHB2 (EPH Receptor B2)* harbours several inactivating mutations in prostate cancer cell lines and clinical prostate cancer samples (Huusko et al. 2004), and a germline nonsense variant in *EPHB2* has been shown to associate with HPC in African American men (Kittles et al. 2006).

The 14.7 kb *EPHA3* deletion has previously been observed to cluster among Finnish hereditary breast and/or ovarian cancer patients (Kuusisto et al. 2013). The authors proposed the deletion to eliminate an intronic regulatory element which, in turn, results in aberrant receptor function. This interpretation is in agreement with the existing data on EPHA3 expression levels in prostate cancer. The up-regulation of the *EPHA3* gene has been detected in androgen-independent prostate cancer cells (Singh et al. 2008). EPHA3 overexpression has also been reported to correlate with a higher Gleason grade in clinical prostate tumour specimens and to contribute to the malignant progression of prostate cancer (Wu et al. 2014a), as well as bone metastasis (Özdemir et al. 2014). Furthermore, in colorectal cancer, high EPHA3 expression correlated significantly with poor survival (Xi & Zhao 2011). Our findings support the hypothesis that germline *EPHA3* deletion contributes to HPC and may be involved in aggressive forms of the disease.

6.3.2 *HOXB3*

The *Homeobox B3 (HOXB3)* gene is located within the same *HOXB* gene cluster at 17q21 as *HOXB13*. Similarly to other *HOX* genes, *HOXB3* is ubiquitously expressed and codes for a TF involved in development (Stelzer et al. 2011). HOXB3 overexpression has been reported to be associated with poor prognosis in acute myeloid leukaemia (Eklund 2011). In addition, the up-regulation of *HOXB3* has been observed in primary prostate cancer tissues. This up-regulation correlates with higher Gleason grades (≥7) and poor survival, suggesting that HOXB3 promotes prostate cancer progression (Chen et al. 2013b). So far, *HOXB3* has not been implicated in genetic predisposition to prostate cancer.

In Study II, two novel *HOXB3* variants, rs10554930 and rs35384813, were identified that were associated with a slightly increased prostate cancer risk but among familial patients only. *In silico* pathogenicity predictors classified both variants as pathogenic. These variants are located in non-coding regions of the genome: rs10554930 approximately 730 bp upstream of the *HOXB3* transcription start site (TSS) and rs35384813 in the 5' untranslated region of the gene. They may be involved in the regulation of *HOXB3* expression, as most regulatory variants have been reported to cluster in promoter regions and near the TSS of the gene that they control (Stranger et al. 2007b). Obviously, functional confirmation is required to support this hypothesis. Considering the importance of *HOX* genes for tissue homeostasis, it is tempting to speculate that *HOXB3* plays a role in prostate cancer susceptibility. Rs10554930 and rs35384813 were, however, detected at a high frequency (>20%) among familial and unselected patients, as well as controls, which makes their pathogenicity less likely. It is possible that these two variants are harmless alone but in combination with other risk variants, participate in modulating the early events that activate the oncogenic process.

6.3.3 *EFCAB13*

The nonsense variant rs118004742 (c.1638T>G, p.Tyr546Ter) in the *EFCAB13* gene was observed to associate weakly (nominal $p = 0.048$) with increased prostate cancer risk among Finnish HPC patients in Study II. The variant was detected in 19 out of 188 families, but complete co-segregation with affection status was recorded for only three families. Among unselected patients, no evidence for association with the disease could be shown.

The *EFCAB13 (EF-Hand Calcium Binding Domain 13)* gene, also known as *C17orf57*, is located at the prostate-cancer-linked locus 17q21.3. A limited amount of data are available for this gene or its protein product. The EFCAB13 protein is predicted to contain a particular structure, an EF-hand, which is the most common calcium-binding motif found in proteins (Lewit-Bentley & Réty 2000). Upon calcium ion binding, the EF-hand motif undergoes a conformational change that results in the activation of the protein. EFCAB13 may thus be involved in the detection and modulation of calcium signals. It is expressed in various tissues, including the prostate, and GO data support nuclear or cytoplasmic localization (Ashburner et al. 2000, Stelzer et al. 2011). According to STRING v10, a database designed to provide illustrations of protein interaction networks, EFCAB13 associates with several class V and class IX myosin proteins (Szklarczyk et al. 2015). Myosins are actin-based motor molecules involved in intracellular movements, vesicular and membrane trafficking and actin cytoskeleton remodelling. The most interesting partner is the myosin VB protein, which participates in epithelial cell polarization (Stelzer et al. 2011).

Rs118004742 creates a premature stop codon in the *EFCAB13* transcript, leading to the production of a severely truncated protein. As expected, the *in silico* predictors classified the variant as pathogenic. The shortened protein may function abnormally, but equally likely, the mRNA molecule containing the premature translation termination codon will undergo nonsense-mediated mRNA decay, and no protein is produced from the defective allele. However, without functional studies, these interpretations remain only speculative. Additional reports uniting *EFCAB13* with prostate cancer have not been published. Therefore, the contribution of this gene to familial prostate cancer awaits further validation, and for the time being, the detected association remains suggestive.

6.4 eQTL variants and prostate cancer risk (II)

The traditional eQTL analysis aimed to identify regulatory SNPs for only those genes that were differentially expressed between patients and controls. The strongest regulatory potential was found for two SNPs, rs12620966 on chromosome 2 and rs11650354 on chromosome 17. Rs12620966 was associated with the differential expression levels of *AGAP1*, which codes for a GTPase-activating protein involved in membrane trafficking and cytoskeleton dynamics (Nie et al. 2002). *AGAP1* has not been associated with prostate cancer, but high expression levels have been reported to correlate with good prognosis in paediatric high-risk B-precursor acute lymphoblastic leukaemia (Harvey et al. 2010). The target gene of rs11650354, *TBKBP1*, encodes an adapter protein that participates in antiviral innate immunity (Stelzer et al. 2011). Rs11650354 is a known eQTL, and its association with *TBKBP1* regulation has been reported previously (Zeller et al. 2010).

The modified eQTL analysis investigated the regulatory role of SNPs with a previously established association with prostate cancer (Eeles et al. 2013), and two interesting *cis*-eQTLs on chromosome 17 were identified. Rs4793943 was observed to regulate the expression of the *ZNF652* gene, providing further evidence that ZNF652 plays a role in prostate carcinogenesis. The second eQTL, rs4793976, was associated with the expression levels of the *SPOP* gene. *SPOP* mutations have been detected in 6-13% of primary prostate cancers negative for the *TMPRSS2-ERG* fusion (Barbieri et al. 2012) and are regarded as driver lesions that define a distinct molecular subclass of prostate cancer (Barbieri et al. 2013). In a study by Zuhlke and colleagues, a germline missense mutation in *SPOP* was observed to segregate completely with affection status in a prostate cancer family, suggesting that *SPOP* may be a candidate gene for HPC (Zuhlke et al. 2014). Another study demonstrated that SPOP modulates DNA double-strand break repair and that *SPOP* mutations are associated with genomic instability (Boysen et al. 2015).

These four regulatory SNPs were found to be located within the non-coding regions of the genome. Rs11650354 resided within the *TBX21 (T-Box 21)* gene, which codes for a TF involved in the regulation of developmental processes (Stelzer et al. 2011). The remaining three SNPs were located within non-coding RNA genes. Transcriptional regulation is known to be an important mechanism underlying cancer predisposition (Monteiro & Freedman 2013). Furthermore, non-coding RNAs have been reported to participate in epigenetic pre- and post-transcriptional gene regulation, as well as chromatin assembly and are involved in cancer initiation, development and progression (Bolton et al. 2014). eQTL results should, however, be interpreted with caution. Due to the large number of tests, some SNP-gene connections may represent random observations rather than true associations. It is also important to distinguish statistical significance from biological significance. Therefore, while support for the role of *ZNF652* and *SPOP* as prostate cancer candidate genes was obtained in the eQTL analysis, these results need to be confirmed in functional studies using prostate cancer cells.

6.5 Limitations of the study

The choice of optimal controls is critical for genetic association studies of common late-onset diseases. In a recent combined review of published autopsy studies, the mean prevalence of indolent, non-progressive prostate cancer was reported to be as high as 59% among men aged >79 years (Bell et al. 2015). Therefore, it seems that the incidence of prostate cancer may be underestimated. Association studies are based on a comparison of allele frequencies between different health status groups, but the existence of incidental prostate cancer complicates the reliable assessment of health status. Men with incidental prostate cancer do not manifest any clinical symptoms and are assigned to the unaffected control group, which may then lead to a misinterpretation of the results. One solution to this problem could be the replacement of the traditional case-control setting with a case-case setting, whereby allele frequencies are compared between indolent and aggressive cases rather than between cases and controls. Obviously, this would require the distinction of indolent from progressive cancers, which is currently challenging due to the modest sensitivity of the available cancer detection methods (Vickers et al. 2014, Bell et al. 2015).

The use of blood donors as population controls has been criticized on the grounds that blood donors differ from the general population by several factors, including their medical history and the medical histories of their parents. This might introduce a bias in the interpretation of the results and lead to spurious disease associations (Golding et al. 2013). It is true that in Finland, male blood donors have not been screened for prostate cancer, and their family history of the disease is also unknown. Therefore, these donors may be affected with prostate cancer later in life or carry variants with reduced penetrance that are associated with the disease. The collection of control samples from people with an assessed medical history is often not feasible for individual research groups, as it is both time-consuming and expensive. While blood donors may not optimally represent the genomic constitution of the general population, they do, however, provide a set of controls that is readily available. To reduce the bias for accuracy, sufficiently large numbers of controls should be analysed.

It is also important to select the most appropriate tissue type for genetic studies, as many associations are highly tissue-specific. In cancer studies, especially those focusing on solid tumours, the regulatory eQTL variants detected in the tissue in which the cancer originates are expected to be more informative than are the variants detected in blood, for example (Freedman et al. 2011). However, transcriptome sequencing and subsequent eQTL analysis in Study II were performed on peripheral blood leukocytes rather than on prostate cancer tissue. The primary reason for this was the unavailability of fresh prostate biopsy samples. Post-mortem material was also considered unsuitable, as we were not focusing on the expression profiles of end-stage disease. Instead, the aim was to investigate early regulatory changes that may trigger prostate carcinogenesis. Recently, Diekstra and colleagues exploited an eQTL analysis performed on whole peripheral blood to identify susceptibility genes for amyotrophic lateral sclerosis, a neurodegenerative disorder leading to progressive muscle weakness due to motor neuron loss. They reported that approximately 50% of eQTLs detected in human brain tissue in previous studies overlapped with their data (Diekstra et al. 2012). Therefore, while blood may not be the optimal tissue for identifying prostate-cancer-associated eQTLs, it can provide a valid starting point for investigating gene expression changes that may predispose a patient to the disease. Obviously, eQTLs observed in blood should be validated in prostate cancer tissue to confirm true disease association.

Current genome-wide analysis methods, such as the whole-genome SNP arrays and NGS applications used in Studies II and III, produce extensive amounts of data on genes that are possibly involved in disease susceptibility. This makes the selection of the most appropriate candidate genes for association analysis critical. The choice of the most relevant genes requires prior knowledge of the mechanisms underlying the disease (Kwon & Goate 2000). However, the pathophysiology of prostate cancer is not yet understood in detail, and novel metabolic pathways will undoubtedly be uncovered. Therefore, the ranking of genes is currently based on thorough literature review, information available in disease databases and the use of targeted *in silico* tools

(Patnala et al. 2013). Even at its best, candidate gene selection represents an "educated guess", and as a result, the experimental validation of irrelevant genes cannot be completely avoided. Increased knowledge on the biochemical basis of prostate cancer, together with the continuous development of bioinformatics tools and computational approaches, will make the future selection of candidate genes easier (Wu et al. 2014b, Zhu et al. 2014).

The choice of the most appropriate mutation detection method depends on sample type (fresh, frozen or FFPE), number of samples that need to be analysed, mutation type (SNPs or large genomic rearrangements, for example) and, naturally, the cost. In this thesis study, sequencing, TaqMan chemistry, the Sequenom MassARRAY System and SNP microarray hybridization were used for variant detection. All of these methods are highly accurate. According to the manufacturers, as well as reports on the performance of different techniques, the estimated mutation detection accuracy for the used methods varies from 99.7% to 99.9% (e.g., Rizzo & Buck 2012, Fedick et al. 2013, Lohmann & Klein 2014). Therefore, genotyping errors most likely result from human error. Mistakes can occur at any step during sample preparation or during data analysis and interpretation. NGS data are especially prone to misinterpretation due to uneven read distribution, which can leave regions of the genome uncovered (Rizzo & Buck 2012). This problem can be overcome by increasing the sequencing depth; high-coverage NGS data are considered accurate and highly reliable (Lohmann & Klein 2014). NGS-based sequencing technologies are under constant development. As the sequencing costs per gene decrease, it is likely that improved NGS applications will replace many of the currently used mutation detection methods.

A limiting step in several risk marker studies is the lack of the functional characterization of the identified variants. Assessing the mechanism by which sequence changes act, however, can be challenging. The effect of coding variants on mRNA or protein stability and function can be deduced from the base change, and in addition, database searches and various *in silico* prediction tools aid in this interpretation (Monteiro & Freedman 2013). Instead, establishing the functional consequences of non-coding variants is more difficult. Approximately 90% of cancer-associated SNPs are located in non-coding regions of the genome, and more than 40% reside in intergenic regions (Pomerantz & Freedman 2011). Non-coding variants have been shown to cluster at DNase I hypersensitive sites, indicative of open chromatin and the presence of regulatory DNA regions (Maurano et al. 2012). It is therefore likely that many of the non-coding variants act as eQTLs and control or modulate the expression of their target genes by disrupting TF recognition or binding sites, by altering allelic chromatin states or by forming regulatory networks (Pomerantz & Freedman 2011, Maurano et al. 2012). The functional effects of non-coding variants can be investigated by eQTL analysis (Monteiro & Freedman 2013). Additional evidence for functionality can be obtained from diverse databases, such as the ENCODE database (ENCODE Project Consortium 2012). However, final confirmation of disease association would require the use of *in vitro* and *in vivo* models. The establishment of these models is technically demanding, time-consuming, expensive, and therefore, in practice, generally unachievable for smaller research laboratories.

6.6 Future directions

Prostate cancer is one of the most extensively studied cancers worldwide. Several genetic alterations associated with the disease have been identified, varying from rare, highly penetrant risk variants with obvious functional consequences to common variants contributing only modestly to disease phenotype. Despite these discoveries, this genetically complex and clinically heterogeneous cancer has remained a medical challenge, and improved tools for screening, diagnostics and treatment are required.

Personalized medicine aims to generate individual risk profiles that can be applied to identify high-risk individuals from the general population (Alvarez-Cubero et al. 2013). In addition, patients who are already affected with prostate cancer can be more precisely assigned to clinically defined, distinct subtypes, and their treatment can be tailored according to disease phenotype (Barbieri & Tomlins 2015, Rubin 2015). Instead of individual disease-associated variants, these personalized risk profiles should be based on collections of several SNPs and CNVs (Pomerantz & Freedman 2013). However, while developing the test panels, ethnic diversity needs to be taken into account. In addition to the remarkable worldwide differences in prostate cancer incidence (Center et al. 2012), regional differences should also be addressed. Finland, for example, represents a well-known genetic isolate, and many inherited diseases that are common in Finland are rare elsewhere. The Finnish gene pool originates from a limited number of founders and has been modified by geographic isolation and genetic drift (Peltonen et al. 1999). As a consequence, prostate-cancer-specific risk factors identified in other populations may not be detected among the Finns and cannot be applied in the prediction of cancer risk. Therefore, population-specific gene panels would probably be most informative in prostate cancer risk assessment (Demichelis & Stanford 2015).

Many of the genes pinpointed in this study code for proteins involved in transcriptional regulation and signal transduction, two important cellular processes that are commonly affected in cancer. HOXB13, ZNF652, HDAC4 and HOXB3 act as transcriptional activators or repressors, whereas EPHA3, PDZD2, EFCAB13 and possibly ANO7 participate in intra- and intercellular signalling. While *HOXB13* would be an obvious choice for the Finnish prostate cancer risk panel, the inclusion of the *ZNF652* and *EPHA3* genes should also be considered. *ZNF652* should be considered because of its highly significant association with the disease, and *EPHA3* because of its suggestive association with aggressive cancer. The role of the other candidate genes in prostate carcinogenesis is not entirely clear and requires further functional and clinical validation but will be an interesting topic for future studies.

Next-generation sequencing technologies enable personalized, genome-based medicine in practice. NGS methods have revolutionized laboratory diagnostics for several genetic disorders, including cancer. Diverse targeted NGS strategies can be applied for the analysis of DNA sequence and copy number variation, even in hundreds of genes, simultaneously and at reasonable costs (Luthra et al. 2015). In Finland, NGS-based cancer diagnostics is becoming increasingly popular. For certain cancers, such as breast, ovarian and colorectal cancers, tumour tissue specimens are already routinely sequenced to detect somatic mutations that may influence treatment decisions. In addition, inherited predisposition to these cancers can be determined from blood samples using NGS-based testing methods. For the time being, these tests are offered for at-risk individuals only, generally members of known cancer families. NGS studies also provide novel information on cancer genomes. For prostate cancer, approximately 75 complete genomes and hundreds of exomes have been published, together with countless reports on gene expression and copy number profiles (reviewed in Barbieri & Tomlins 2015).

The first diagnostic panels containing prostate-cancer-associated SNPs were considered to be of limited clinical utility, and in the USA, their clinical use was not recommended (Pomerantz & Freedman 2013). However, the use of NGS technologies in cancer research and diagnostics will continue to increase, and as more information on the genetic alterations predisposing to prostate cancer accumulates, the preclinical identification of asymptomatic at-risk individuals will become feasible. Some European laboratories already offer molecular genetic testing for susceptibility to familial prostate cancer (Table 1). In Finland, these tests are not yet available, mainly because the practical relevance of the results is still restricted. Knowledge of increased cancer risk can also enhance unnecessary anxiety among tested individuals and their family members (Hamilton et al. 2014). The drawbacks and advantages of predictive prostate cancer testing should therefore be carefully considered prior to developing these tests. Genetic counselling protocols also need to be designed to ensure sufficient social and psychological support, in addition to the correct interpretation of test results (Kajula et al. 2015). Predictive testing would most likely benefit members of prostate cancer families, men with elevated PSA levels and individuals who are concerned of their own cancer risk. Therefore, the idea of incorporating prostate cancer genetics in the clinic is definitely worth revisiting in the next few years.

7 Summary and Conclusions

The present study was conducted to obtain new information on genetic factors that predispose individuals to prostate cancer in Finland. While the study focused primarily on the hereditary form of the disease, the findings may aid in the interpretation and understanding of mechanisms underlying sporadic prostate cancer as well.

Two susceptibility loci, 2q37 and 17q11.2-q22, were linked to prostate cancer more than a decade ago, but the identification of causative genes within these loci has proven to be a lingering process. Finally, in 2012, the *HOXB13* gene was mapped to 17q21.3. In this thesis study, the frequency of the *HOXB13* risk variant p.G84E was determined among Finnish familial and unselected prostate cancer patients, and the variant was observed to be exceptionally common among both groups. The variant associated strongly with increased prostate cancer risk, making *HOXB13* the major prostate cancer risk gene in Finland.

The fine-mapping of the 2q37 and 17q11.2-q22 loci by next-generation sequencing and the functional analysis by eQTL mapping resulted in the identification of several prostate-cancer-associated sequence variants in previously reported as well as novel candidate genes. In particular, the role of *ZNF652* as a prostate cancer candidate gene gained additional support, as novel variants contributing significantly to increased cancer risk were identified in this gene. Suggestive evidence for association with hereditary prostate cancer was obtained for the *HDAC4*, *EFCAB13* and *ANO7* genes, but their relevance in prostate cancer predisposition needs to be ascertained in future studies.

Knowledge of the importance of germline copy number changes in human cancer is accumulating rapidly, and further evidence for the involvement of these changes in prostate cancer was obtained in this study. A genome-wide CNV analysis revealed a deletion in the *EPHA3* gene at 3p11.1 that was enriched among Finnish HPC patients, and suggestive association with aggressive disease was discovered.

8 Acknowledgements

This thesis study was carried out in the Laboratory of Cancer Genetics, BioMediTech, University of Tampere and Fimlab Laboratories during the years 2010-2016.

The dean of BioMediTech (BMT), Hannu Hanhijärvi, DDS, PhD and the former medical director of Fimlab Laboratories, docent Erkki Seppälä, MD, PhD are acknowledged for providing excellent research facilities. The BMT's Doctoral Programme in Biomedicine and Biotechnology (TGPBB) is thanked for interesting, multidisciplinary courses and seminars as well as travel grants.

I am grateful to my thesis supervisors, professor Johanna Schleutker, CLG, PhD and docent Tiina Wahlfors, PhD for guiding me into the fascinating, yet challenging world of cancer genetics. Although knowing that I was not heading for a career in science, Johanna unhesitatingly welcomed me in her research group and provided me with an intriguing research topic together with excellent laboratory facilities. Tiina's expertise on prostate cancer genetics, as well as tireless help and support in numerous scientific and practical problems (often involving computers) enabled me to successfully complete this work.

I wish to express my gratitude to my thesis committee members, professor Markus Perola, MD, PhD and professor Mauno Vihinen, PhD for their help during the years.

I sincerely thank the official reviewers of the thesis manuscript, docent Markku Vaarala, MD, PhD and docent Elisabeth Widén, MD, PhD for their constructive criticism, beneficial comments and feedback.

All of the co-authors of my publications are appreciated for their valuable contribution and professional expertise. Specifically, I would like to acknowledge professor Teuvo L.J. Tammela, MD, PhD, Leena Saaristo, MD, Elisa Vuorinen, MSc and "the statistics guys" Tommi Rantapero, MSc, Daniel Fischer, MSc and Oyediran "Prince" Akinrinade, PhD.

I wish to thank all the members of our research group, Genetic Predisposition to Cancer. Especially, I want to thank my closest colleagues Riikka Nurminen, MSc, Kirsi Määttä, PhD, Sanna Siltanen, PhD, Ms. Riina Kylätie and the TGPBB coordinator Henna Mattila, PhD for sharing the ups and downs of scientific as well as personal life. It has been a privilege to get to know you all! Additionally, I wish to express my gratitude to Linda Enroth, MSc, Ms. Riitta Vaalavuo, Ms. Kirsi Rouhento and other co-workers and friends at BioMediTech. BMT's IT designer Jukka Lehtiniemi is especially acknowledged for his amazing image processing skills.

I also want to thank my former and present colleagues in the Laboratory of Clinical Genetics, Fimlab Laboratories. In particular, I wish to acknowledge docent Ritva Karhu, CLG, PhD for offering me the opportunity to take time off from work and finish writing the thesis manuscript. Tiina Lund-Aho, CLG, PhD and Kirsi Kaukoniemi, CLG (it.), MSc are thanked for their loyal, although somewhat noisy friendship and unconditional support. Leena Huhti, CLG (it.), PhD, docent Marketta Kähkönen, CLG, PhD, Kyllikki Haapala, CLG, MSc as well as other co-workers and friends at Fimlab are also warmly acknowledged.

In addition, I wish to thank Juha Pursiheimo, PhD for his kind help during the preparation of the thesis manuscript.

Most importantly, I want to thank my closest ones, my parents Pirkko-Liisa and Aarne, my brother Marko, my husband Timo and my lovely daughters Pinja, Eerika and Henna. Over the years, you have given me the time and space I have needed to complete this project, and I truly appreciate your efforts. You have always stood by me, no matter what, and shared the moments of despair and joy with me. Without your love, encouragement and unwavering faith I would have never been able to cross the finishing line.

Lastly, I want to express my sincere gratitude to the cancer patients and their family members for participating in this study.

This work was financially supported by the University of Tampere, the Tampere Graduate Program in Biomedicine and Biotechnology (TGPBB), the Competitive Research Funding of the Tampere University Hospital, the Helsinki University Hospital Research Funds, the Academy of Finland, the Finnish Cancer Organizations, the Cancer Society of Finland, the Sigrid Juselius Foundation, the Ida Montin Foundation, the Education Fund and the Scientific Foundation of the City of Tampere.

Tampere, May 2016

Virpi Laitinen

9 References

Aaltonen LA, Salovaara R, Kristo P, et al. (1998). Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *New Engl J Med* 338:1481-1487.

Adamczak R, Porollo A & Meller J (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 59:467-475.

Adzhubei I, Jordan DM & Sunyaev SR (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7:Unit7.20.

Akbari MR, Trachtenberg J, Lee J, et al. (2012). Association between germline HOXB13 G84E mutation and risk of prostate cancer. *J Natl Cancer Inst* 104:1260-1262.

Allott EH, Masko EM & Freedland SJ (2013). Obesity and prostate cancer: weighing the evidence. *Eur Urol* 63:800-809.

Almal SH & Padh H (2012). Implications of gene copy-number variation in health and diseases. *J Hum Genet* 57:6-13.

Alvarez-Cubero MJ, Saiz M, Martinez-Gonzalez LJ, et al. (2013). Genetic analysis of the principal genes related to prostate cancer: a review. *Urol Oncol* 31:1419-1429.

Amin Al Olama A, Dadaev T, Hazelett DJ, et al. (2015). Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. *Hum Mol Genet* 24:5589-5602.

Amundadottir LT, Sulem P, Gudmundsson J, et al. (2006). A common variant associated with prostate cancer in European and African populations. *Nat Genet* 38:652-658.

Anders S & Huber W (2010). Differential expression analysis for sequence count data. *Genome Biol* 11:R106.

Anders S, Pyl PT & Huber W (2015). HTseq – a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166-169.

Ashburner M, Ball CA, Blake JA, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25-29.

Attard G, Parker C, Eeles RA, et al. (2016). Prostate cancer. *Lancet* 387:70-82.

Autran-Gomez AM, Scarpa RM & Chin J (2012). High-intensity focused ultrasound and cryotherapy as salvage treatment in local radio-recurrent prostate cancer. *Urol Int* 89:373-379.

Aziz N, Zhao Q, Bry L, et al. (2015). College of American pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med* 139:481-493.

Bailey-Wilson JE, Childs EJ, Cropp CD, et al. (2012). Analysis of Xq27-28 linkage in the International Consortium for Prostate Cancer Genetics (ICPCG) families. *BMC Med Genet* 13:46.

Barbieri CE & Tomlins SA (2015). Reprint of: the prostate cancer genome: perspectives and potential. *Urol Oncol* 33:95-102.

Barbieri CE, Baca SC, Lawrence MS, et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 44:685-689.

Barbieri CE, Bangma CH, Bjartell A, et al. (2013). The mutational landscape of prostate cancer. *Eur Urol* 64:567-576.

Barrett JC, Fry B, Maller J, et al. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263-265.

Bell KJL, Del Mar C, Wright G, et al. (2015). Prevalence of incidental prostate cancer: a systematic review of autopsy studies. *Int J Cancer* 137:1749-1757.

Bera TK, Das S, Maeda H, et al. (2004). NGEP, a gene encoding a membrane protein detected only in prostate cancer and normal prostate. *Proc Natl Acad Sci USA* 101:3059-3064.

Bhatlekar S, Fields JZ & Boman BM (2014). HOX genes and their role in the development of human cancers. *J Mol Med* 92:811-823.

Bhavsar A & Verma S (2014). Anatomic imaging of the prostate. *Biomed Res Int* 2014:728539.

Bolton EM, Tuzova AV, Walsh AL, et al. (2014). Noncoding RNAs in prostate cancer: the long and the short of it. *Clin Cancer Res* 20:35-43.

Boniol M, Autier P, Perrin P, et al. (2015). Variation of prostate-specific antigen value in men and risk of high-grade prostate cancer: analysis of the prostate, lung, colorectal, and ovarian cancer screening trial study. *Urology* 85:1117-1122.

Bostwick DG & Cheng L (2012). Precursors of prostate cancer. *Histopathology* 60:4-27.

Bostwick DG, Shan A, Qian J, et al. (1998). Independent origin of multiple foci of prostatic intraepithelial neoplasia: comparison with matched foci of prostate carcinoma. *Cancer* 83:1995-2002.

Bostwick DG, Burke HB, Djakiew D, et al. (2004). Human prostate cancer risk factors. *Cancer* 101:2371-2490.

Boyle AP, Hong EL, Hariharan M, et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22:1790-1797.

Boysen G, Barbieri CE, Prandi D, et al. (2015). SPOP mutation leads to genomic instability in prostate cancer. *eLife* 4:e09207.

Brandt A, Bermejo JL, Sundquist J, et al. (2010). Age-specific risk of incident prostate cancer and risk of death from prostate cancer defined by the number of affected family members. *Eur Urol* 58:275-280.

Breyer JP, Avritt TG, McReynolds KM, et al. (2012). Confirmation of the HOXB13 G84E germline mutation in familial prostate cancer. *Cancer Epidemiol Biomarkers Prev* 21:1348-1353.

Bromberg Y & Rost B (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* 35:3823-3835.

Burkhart DL & Sage J (2008). Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat Rev Cancer* 8:671-682.

Cairns P, Okami K, Halachmi S, et al. (1997). Frequent inactivation of PTEN/MMAC1 in primary prostate cancer. *Cancer Res* 57:4997-5000.

Calabrese R, Capriotti E, Fariselli P, et al. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* 30:1237-1244.

Callen DF, Ricciardelli C, Butler M, et al. (2010). Co-expression of the androgen receptor and the transcription factor ZNF652 is related to prostate cancer outcome. *Oncol Rep* 23:1045-1052.

Capriotti E, Calabrese R & Casadio R (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22:2729-2734.

Capriotti E, Fariselli P, Rossi I, et al. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* 9:S6.

Carpten J, Nupponen N, Isaacs S, et al. (2002). Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. *Nat Genet* 30:181-184.

Carter BS, Beaty TH, Steinberg GD, et al. (1992). Mendelian inheritance of familial prostate cancer. *Proc Natl Acad Sci USA* 89:3367-3371.

Carter BS, Bova GS, Beaty TH, et al. (1993). Hereditary prostate cancer: epidemiologic and clinical features. *J Urol* 150:797-802.

Center MM, Jemal A, Lortet-Tieulent J, et al. (2012). International variation in prostate cancer incidence and mortality rates. *Eur Urol* 61:1079-1092.

Cerami EG, Gross BE, Demir E, et al. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* 39:D685-D690.

Chen Z & Lu W (2015). Roles of ubiquitination and SUMOylation on prostate cancer: mechanisms and clinical implications. *Int J Mol Sci* 16:4560-4580.

Chen Z, Greenwood C, Isaacs WB, et al. (2013a). The G84E mutation of HOXB13 is associated with increased risk for prostate cancer: results from the REDUCE trial. *Carcinogenesis* 34:1260-1264.

Chen J, Zhu S, Jiang N, et al. (2013b). HoxB3 promotes prostate cancer cell progression by transactivating CDCA3. *Cancer Lett* 330:217-224.

Cheng J, Randall A & Baldi P (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 62:1125-1132.

Cheng L, Montironi R, Bostwick DG, et al. (2012). Staging of prostate cancer. *Histopathology* 60:87-117.

Choi Y, Sims GE, Murphy S, et al. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 7:e46688.

Clark J, Merson S, Jhavar S, et al. (2007). Diversity of TMPRSS2-ERG fusion transcripts in the human prostate. *Oncogene* 26:2667-2673.

Cohen AL, Piccolo SR, Cheng L, et al. (2013). Genomic pathway analysis reveals that EZH2 and HDAC4 represent mutually exclusive epigenetic pathways across human cancers. *BMC Med Genomics* 6:35.

Cole DE, Gallinger S, McCready DR, et al. (1996). Genetic counselling and testing for susceptibility to breast, ovarian and colon cancer: where are we today? *Can Med Assoc J* 154:149-155.

Conrad DF, Pinto D, Redon R, et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464:704-712.

Cooper GM & Shendure J (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 12:628-640.

Crawford ED (2003). Epidemiology of prostate cancer. *Urology* 62:3-12.

Cropp CD, Simpson CL, Wahlfors T, et al. (2011). Genome-wide linkage scan for prostate cancer susceptibility in Finland: evidence for a novel locus on 2q37.3 and confirmation of signal on 17q21-q22. *Int J Cancer* 129:2400-2407.

Cussenot O, Valeri A, Berthon P, et al. (1998). Hereditary prostate cancer and other genetic predispositions to prostate cancer. *Urol Int* 60:30-34.

Damaschke NA, Yang B, Bhusari S, et al. (2013). Epigenetic susceptibility factors for prostate cancer with aging. *Prostate* 73:1721-1730.

Das S, Hahn Y, Nagata S, et al. (2007). NGEP, a prostate-specific plasma membrane protein that promotes the association of LNCaP cells. *Cancer Res* 67:1594-1601.

Das S, Hahn Y, Walker DA, et al. (2008). Topology of NGEP, a prostate specific cell:cell junction protein widely expressed in many cancers of different grade level. *Cancer Res* 68:6306-6312.

Demichelis F & Stanford JL (2015). Genetic predisposition to prostate cancer: update and future perspectives. *Urol Oncol* 33:75-84.

Demichelis F, Fall K, Perner S, et al. (2007). TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene* 26:4596-4599.

Demichelis F, Setlur SR, Banerjee S, et al. (2012). Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proc Natl Acad Sci USA* 109:6686-6691.

Diekstra FP, Saris CGJ, van Rheenen W, et al. (2012). Mapping of gene expression reveals CYP27A1 as a susceptibility gene for sporadic ALS. *PLoS ONE* 7:e35333.

Diskin SJ, Hou C, Glessner JT, et al. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459:987-991.

Dobosy JR, Roberts JLW, Fu VX, et al. (2007). The expanding role of epigenetics in the development, diagnosis and treatment of prostate cancer and benign prostatic hyperplasia. *J Urol* 177:822-831.

Dong X, Wang L, Taniguchi K, et al. (2003). Mutations in CHEK2 associated with prostate cancer risk. *Am J Hum Genet* 72:270-280.

Duran C, Qu Z, Osunkoya AO, et al. (2012). ANOs 3-7 in the anoctamin/Tmem16 Cl$^-$ channel family are intracellular proteins. *Am J Physiol Cell Physiol* 302:C482-C493.

Edwards SM, Kote-Jarai Z, Meitz J, et al. (2003). Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *Am J Hum Genet* 72:1-12.

Eeles RA, Kote-Jarai Z, Giles GG, et al. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* 40:316-321.

Eeles RA, Amin Al Olama A, Benlloch S, et al. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* 45:385-391.

Eeles R, Goh C, Castro E, et al. (2014). The genetic epidemiology of prostate cancer and its clinical implications. *Nat Rev Urol* 11:18-31.

Eerola H, Blomqvist C, Pukkala E, et al. (2000). Familial breast cancer in southern Finland: how prevalent are breast cancer families and can we trust the family history reported by patients? *Eur J Cancer* 36:1143-1148.

Eklund E (2011). The role of Hox proteins in leukemogenesis: insights into key regulatory events in hematopoiesis. *Crit Rev Oncog* 16:65-76.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-74.

Epstein JI, Egevad L, Amin MB, et al. (2016). The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol* 40:244-252.

Erkko H, Xia B, Nikkilä J, et al. (2007). A recurrent mutation in PALB2 in Finnish cancer families. *Nature* 446:316-319.

Ewing CM, Ray AM, Lange EM, et al. (2012). Germline mutations in HOXB13 and prostate-cancer risk. *New Engl J Med* 366:141-149.

Fedewa SA & Jemal A (2013). Prostate cancer disease severity and country of origin among black men in the United States. *Prostate Cancer P D* 16:176-180.

Fedick A, Su J, Jalas C, et al. (2013). High-throughput carrier screening using TaqMan allelic discrimination. *PLoS ONE* 8:e59722.

Feuk L, Carson AR & Scherer SW (2006). Structural variation in the human genome. *Nat Rev Genet* 7:85-97.

Finnish Cancer Registry, www.cancerregistry.fi, updated on 05.03.2016.

Firth HV, Richards SM, Bevan AP, et al. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet* 84:524-533.

Fischer D, Oja H, Schleutker J, et al. (2014). Generalized Mann-Whitney type tests for microarray experiments. *Scand J Stat* 41:672-692.

Fischle W, Kiermer V, Dequiedt F, et al. (2001). The emerging role of class II histone deacetylases. *Biochem Cell Biol* 79:337-348.

Fitzgerald LM, Kumar A, Boyle EA, et al. (2013). Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* 22:1520-1528.

Flicek P, Ahmed I, Amode MR, et al. (2013). Ensembl 2013. *Nucleic Acids Res* 41:D48-D55.

Forbes SA, Beare D, Gunasekaran P, et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43:D805-D811.

Foulkes WD (2008). Inherited susceptibility to common cancers. *New Engl J Med* 359:2143-2153.

Fox BP, Tabone CJ & Kandpal RP (2006). Potential clinical relevance of Eph receptors and ephrin ligands expressed in prostate carcinoma cell lines. *Biochem Biophys Res Commun* 342:1263-1272.

Fraser M, Berlin A, Bristow RG, et al. (2015). Genomic, pathological, and clinical heterogeneity as drivers of personalized medicine in prostate cancer. *Urol Oncol* 33:85-94.

Freedman ML, Monteiro ANA, Gayther SA, et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* 43:513-518.

Gillanders EM, Xu J, Chang BL, et al. (2004). Combined genome-wide scan for prostate cancer susceptibility genes. *J Natl Cancer Inst* 96:1240-1247.

GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. http://globocan.iarc.fr, accessed on 23.10.2015.

Golding J, Northstone K, Miller LL, et al. (2013). Differences between blood donors and a population sample: implications for case-control studies. *Int J Epidemiol* 42:1145-1156.

Grandori C, Cowley SM, James LP, et al. (2000). The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* 16:653-699.

Greene KL, Albertsen PC, Babaian RJ, et al. (2013). Prostate specific antigen best practice statement: 2009 update. *J Urol* 189:S2-S11.

Grönberg H, Damber L & Damber JE (1996). Familial prostate cancer in Sweden. A nationwide register cohort study. *Cancer* 77:138-143.

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

Gudmundsson J, Sulem P, Steinthorsdottir V, et al. (2007b). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* 39:977-983.

Gundem G, Van Loo P, Kremeyer B, et al. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature* 520:353-357.

Göring HH, Curran JE, Johnson MP, et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 39:1208-1216.

Hackshaw-McGeagh LE, Perry RE, Leach VA, et al. (2015). A systematic review of dietary, nutritional, and physical activity interventions for the prevention of prostate cancer progression and mortality. *Cancer Cause Control* 26:1521-1550.

Haiman CA, Le Marchand L, Yamamoto J, et al. (2007). A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* 39:954-956.

Haiman CA, Chen GK, Blot WJ, et al. (2011). Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat Genet* 43:570-573.

Halkidou K, Cook S, Leung HY, et al. (2004). Nuclear accumulation of histone deacetylase 4 (HDAC4) coincides with the loss of androgen sensitivity in hormone refractory cancer of the prostate. *Eur Urol* 45:382-389.

Hamilton JG, Edwards HM, Khoury MJ, et al. (2014). Cancer screening and genetics: a tale of two paradigms. *Cancer Epidemiol Biomarkers Prev* 23:909-916.

Han Y, Hazelett DJ, Wiklund F, et al. (2015). Integration of multiethnic fine-mapping and genomic annotation to prioritize candidate functional SNPs at prostate cancer susceptibility regions. *Hum Mol Genet* 24:5603-5618.

Hanahan D & Weinberg RA (2011). Hallmarks of cancer: the next generation. *Cell* 144:646-674.

Harvey RC, Mullighan CG, Wang X, et al. (2010). Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood* 116:4874-4884.

Hemminki K (2012). Familial risk and familial survival in prostate cancer. *World J Urol* 30:143-148.

Hemminki K & Czene K (2002). Age specific and attributable risks of familial prostate carcinoma from the family-cancer database. *Cancer* 95:1346-1353.

Hjelmborg JB, Scheike T, Holst K, et al. (2014). The heritability of prostate cancer in the Nordic twin study of cancer. *Cancer Epidemiol Biomarkers Prev* 23:2303-2310.

Hori S, Butler E & McLoughlin J (2011). Prostate cancer and diet: food for thought? *BJU Int* 107:1348-1359.

Horne SD, Pollick SA & Heng HHQ (2015). Evolutionary mechanism unifies the hallmarks of cancer. *Int J Cancer* 136:2012-2021.

Huang L, Pu Y, Hepps D, et al. (2007a). Posterior Hox gene expression and differential androgen regulation in the developing and adult rat prostate lobes. *Endocrinology* 148:1235-1245.

Huang LT, Gromiha MM & Ho SY (2007b). iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* 23:1292-1293.

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

Huusko P, Ponciano-Jackson D, Wolf M, et al. (2004). Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer. *Nat Genet* 36:979-983.

Ikonen T, Matikainen MP, Syrjäkoski K, et al. (2003). BRCA1 and BRCA2 mutations have no major role in predisposition to prostate cancer in Finland. *J Med Genet* 40:e98.

Isaacs W & Kainu T (2001). Oncogenes and tumor suppressor genes in prostate cancer. *Epidemiol Rev* 23:36-41.

Itsara A, Cooper GM, Baker C, et al. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84:148-161.

James ND, Spears MR, Clarke NW, et al. (2015). Survival with newly diagnosed metastatic prostate cancer in the "Docetaxel era": data from 917 patients in the control arm of the STAMPEDE trial (MRC PR08, CRUK/06/019). *Eur Urol* 67:1028-1038.

Jin G, Sun J, Liu W, et al. (2011). Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. *Carcinogenesis* 32:1057-1062.

Jin G, Lu L, Cooney KA, et al. (2012). Validation of prostate cancer risk-related loci identified from genome-wide association studies using family-based association analysis: evidence from the International Consortium for Prostate Cancer Genetics (ICPCG). *Hum Genet* 131:1095-1103.

Johnson AM, Zuhlke KA, Plotts C, et al. (2014). Mutational landscape of candidate genes in familial prostate cancer. *Prostate* 74:1371-1378.

Kajula O, Kääriäinen M, Moilanen JS, et al. (2015). The quality of genetic counseling and connected factors as evaluated by male BRCA1/2 mutation carriers in Finland. *J Genet Couns* [Epub ahead of print].

Kanehisa M & Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28:27-30.

Karlsson R, Aly M, Clements M, et al. (2014). A population-based assessment of germline HOXB13 G84E mutation and prostate cancer risk. *Eur Urol* 65:169-176.

Khemlina G, Ikeda S & Kurzrock R (2015). Molecular landscape of prostate cancer: Implications for current clinical trials. *Cancer Treat Rev* 41:761-766.

Kicinski M, Vangronsveld J & Nawrot TS (2011). An epidemiological reappraisal of the familial aggregation of prostate cancer: a meta-analysis. *PLoS ONE* 6:e27130.

Kilpeläinen TP, Auvinen A, Määttänen L, et al. (2010). Results of the three rounds of the Finnish prostate cancer screening trial – the incidence of advanced cancer is decreased by screening. *Int J Cancer* 127:1699-1705.

Kim YR, Oh KJ, Park RY, et al. (2010). HOXB13 promotes androgen independent growth of LNCaP prostate cancer cells by the activation of E2F signaling. *Mol Cancer* 9:124.

Kim IJ, Kang TW, Jeong T, et al. (2014). HOXB13 regulates the prostate-derived Ets factor: implications for prostate cancer cell invasion. *Int J Oncol* 45:869-876.

Kittles RA, Baffoe-Bonnie AB, Moses TY, et al. (2006). A common nonsense mutation in EphB2 is associated with prostate cancer risk in African American men with a positive family history. *J Med Genet* 43:507-511.

Klug A (2010). The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* 79:213-231.

Knudson AG (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 68:820-823.

Kochetkova M, McKenzie OLD, Bais AJ, et al. (2002). CBFA2T3 (MTG16) is a putative breast tumor suppressor gene from the breast cancer loss of heterozygosity region at 16q24.3. *Cancer Res* 62:4599-4604.

Kote-Jarai Z, Amin Al Olama A, Giles GG, et al. (2011). Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet* 43:785-791.

Kouprina N, Pavlicek A, Noskov VN, et al. (2005). Dynamic structure of the SPANX gene cluster mapped to the prostate cancer susceptibility locus HPCX at Xq27. *Genome Res* 15:1477-1486.

Kouprina N, Noskov VN, Solomon G, et al. (2007). Mutational analysis of SPANX genes in families with X-linked prostate cancer. *Prostate* 67:820-828.

Krepischi ACV, Pearson PL & Rosenberg C (2012a). Germline copy number variations and cancer predisposition. *Future Oncol* 8:441-450.

Krepischi ACV, Achatz MIW, Santos EMM, et al. (2012b). Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* 14:R24.

Kuiper RP, Ligtenberg MJL, Hoogerbrugge N, et al. (2010). Germline copy number variation and cancer risk. *Curr Opin Genet Dev* 20:282-289.

Kumar R, Manning J, Spendlove HE, et al. (2006). ZNF652, a novel zinc finger protein, interacts with the putative breast tumor suppressor CBFA2T3 to repress transcription. *Mol Cancer Res* 4:655-665.

Kumar R, Cheney KM, McKirdy R, et al. (2008). CBFA2T3-ZNF652 corepressor complex regulates transcription of the E-box gene HEB. *J Biol Chem* 283:19026-19038.

Kumar P, Henikoff S & Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073-1081.

Kumar R, Selth LA, Schultz RB, et al. (2011). Genome-wide mapping of ZNF652 promoter binding sites in breast cancer cells. *J Cell Biochem* 112:2742-2747.

Kung JTY, Colognori D & Lee JT (2013). Long noncoding RNAs: past, present, and future. *Genetics* 193:651-669.

Kuusisto KM, Bebel A, Vihinen M, et al. (2011). Screening for BRCA1, BRCA2, CHEK2, PALB2, BRIP1, RAD50, and CDH1 mutations in high-risk Finnish BRCA1/2-founder mutation-negative breast and/or ovarian cancer individuals. *Breast Cancer Res* 13:R20.

Kuusisto KM, Akinrinade O, Vihinen M, et al. (2013). Copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. *PLoS ONE* 8:e71802.

Kwon JM & Goate AM (2000). The candidate gene approach. *Alcohol Res Health* 24:164-168.

Laity JH, Lee BM & Wright PE (2001). Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struc Biol* 11:39-46.

Lange EM, Gillanders EM, Davis CC, et al. (2003). Genome-wide scan for prostate cancer susceptibility genes using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1. *Prostate* 57:326-334.

Lange EM, Robbins CM, Gillanders EM, et al. (2007). Fine-mapping the putative chromosome 17q21-22 prostate cancer susceptibility gene to a 10 cM region based on linkage analysis. *Hum Genet* 121:49-55.

Lappalainen T, Montgomery SB, Nica AC, et al. (2011). Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet* 89:459-463.

Lappalainen T, Sammeth M, Friedländer MC, et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506-511.

Larson NB, McDonnell S, French AJ, et al. (2015). Comprehensively evaluating cis-regulatory variation in the human prostate transcriptome by using gene-level allele-specific expression. *Am J Hum Genet* 96:869-882.

Ledet EM, Hu X, Sartor O, et al. (2013). Characterization of germline copy number variation in high-risk African American families with prostate cancer. *Prostate* 73:614-623.

Leongamornlert D, Saunders E, Dadaev T, et al. (2014). Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease. *Brit J Cancer* 110:1663-1672.

Levine AJ (1990). Tumor suppressor genes. *Bioessays* 12:60-66.

Lewit-Bentley A & Réty S (2000). EF-hand calcium-binding proteins. *Curr Opin Struct Biol* 10:637-643.

Li Q, Stram A, Chen C, et al. (2014). Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Hum Mol Genet* 23:5294-5302.

Li Y, Wang X, Vural S, et al. (2015). Exome analysis reveals differentially mutated gene signatures of stage, grade and subtype in breast cancers. *PLoS ONE* 10:e0119383.

Lichtenstein P, Holm NV, Verkasalo PK, et al. (2000). Environmental and heritable factors in the causation of cancer. Analyses of cohorts of twins from Sweden, Denmark, and Finland. *New Engl J Med* 343:78-85.

Lilja H, Ulmert D, Björk T, et al. (2007). Long-term prediction of prostate cancer up to 25 years before diagnosis of prostate cancer using prostate kallikreins measured at age 44 to 50 years. *J Clin Oncol* 25:431-436.

Lin X, Qu L, Chen Z, et al. (2013). A novel germline mutation in HOXB13 is associated with prostate cancer risk in Chinese men. *Prostate* 73:169-175.

Lin PH, Aronson W & Freedland SJ (2015). Nutrition, dietary interventions and prostate cancer: the latest evidence. *BMC Med* 13:3.

Lindström S, Schumacher FR, Cox D, et al. (2012). Common genetic variants in prostate cancer risk prediction – Results from the NCI breast and prostate cancer cohort consortium (BPC3). *Cancer Epidemiol Biomarkers Prev* 21:437-444.

Linja MJ & Visakorpi T (2004). Alterations of androgen receptor in prostate cancer. *J Steroid Biochem Mol Biol* 92:255-264.

Lisabeth EM, Fernandez C & Pasquale EB (2012). Cancer somatic mutations disrupt functions of the EphA3 receptor tyrosine kinase through multiple mechanisms. *Biochemistry* 51:1464-1475.

Liu W, Laitinen S, Khan S, et al. (2009a). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* 15:559-565.

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

Liu YP, Hu FL, Li DD, et al. (2011). Does physical activity reduce the risk of prostate cancer? A systematic review and meta-analysis. *Eur Urol* 60:1029-1044.

Lloyd T, Hounsome L, Mehay A, et al. (2015). Lifetime risk of being diagnosed with, or dying from, prostate cancer by major ethnic group in England 2008-2010. *BMC Med* 13:171.

Loeb S, Carter HB, Catalona WJ, et al. (2012). Baseline prostate-specific antigen testing at a young age. *Eur Urol* 61:1-7.

Lohmann K & Klein C (2014). Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics* 11:699-707.

Luthra R, Chen H, Roy-Chowdhuri S, et al. (2015). Next-generation sequencing in clinical molecular diagnostics of cancer: advantages and challenges. *Cancers* 7:2023-2036.

Lynch HT & Shaw TG (2013). Familial prostate cancer and HOXB13 founder mutations: geographic and racial/ethnic variations. *Hum Genet* 132:1-4.

MacArthur DG, Manolio TA, Dimmock DP, et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469-476.

MacDonald JR, Ziman R, Yuen RK, et al. (2014). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 42:D986-D992.

Macintosh CA, Stower M, Reid N, et al. (1998). Precise microdissection of human prostate cancers reveals genotypic heterogeneity. *Cancer Res* 58:23-28.

Majewski J & Pastinen T (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* 27: 72-79.

Maqungo M, Kaur M, Kwofie SK, et al. (2011). DDPC: Dragon Database of genes associated with Prostate Cancer. *Nucleic Acids Res* 39:D980-D985.

Mardis ER (2008). Next-generation DNA sequencing methods. *Annu Rev Genom Hum Genet* 9:387-402.

Mardis ER & Wilson RK (2009). Cancer genome sequencing: a review. *Hum Mol Genet* 18:R163-R168.

Maurano MT, Humbert R, Rynes E, et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190-1195.

Metzker ML (2010). Sequencing technologies – the next generation. *Nat Rev Genet* 11:31-46.

Meyer D, Zeileis A & Hornik K (2014). VCD: Visualizing Categorical Data. R package version 1.3-2.

Mi H, Dong Q, Muruganujan A, et al. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res* 38:D204-D210.

Michaelson JJ, Loguercio S & Beyer A (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 48:265-276.

Mohsenzadegan M, Madjd Z, Asgari M, et al. (2013). Reduced expression of NGEP is associated with high-grade prostate cancers: a tissue microarray analysis. *Cancer Immunol Immunother* 62:1609-1618.

Moir-Meyer GL, Pearson JF, Lose F, et al. (2015). Rare germline copy number deletions of likely functional importance are implicated in endometrial cancer predisposition. *Hum Genet* 134:269-278.

Monteiro ANA & Freedman ML (2013). Lessons from postgenome-wide association studies: functional analysis of cancer predisposition loci. *J Intern Med* 274:414-424.

Nelson WG, De Marzo AM & Yegnasubramanian S (2014). The diet as a cause of human prostate cancer. *Cancer Treat Res* 159:51-68.

Nica AC & Dermitzakis ET (2013). Expression quantitative trait loci: present and future. *Phil Trans R Soc B* 368:20120362.

Nie Z, Stanley KT, Stauffer S, et al. (2002). AGAP1, an endosome-associated, phosphoinositide-dependent ADP-ribosylation factor GTPase-activating protein that affects actin cytoskeleton. *J Biol Chem* 277:48965-48975.

Norris JD, Chang CY, Wittmann BM, et al. (2009). The homeodomain protein HOXB13 regulates the cellular response to androgens. *Mol Cell* 36:405-416.

Nurminen R, Lehtonen R, Auvinen A, et al. (2013). Fine mapping of 11q13.5 identifies regions associated with prostate cancer and prostate cancer death. *Eur J Cancer* 49:3335-3343.

Oesterling JE, Cooner WH, Jacobsen SJ, et al. (1993). Influence of patient age on the serum PSA concentration. An important clinical observation. *Urol Clin North Am* 20:671-680.

Olatubosun A, Väliaho J, Härkönen J, et al. (2012). PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat* 33:1166-1174.

Orsted DD, Bojesen SE, Nielsen SF, et al. (2011). Association of clinical benign prostate hyperplasia with prostate cancer incidence and mortality revisited: a nationwide cohort study of 3,009,258 men. *Eur Urol* 60:691-698.

Ozsolak F & Milos PM (2011). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 12:87-98.

Pakkanen S, Wahlfors T, Siltanen S, et al. (2009). PALB2 variants in hereditary and unselected Finnish prostate cancer cases. *J Negat Results Biomed* 8:12.

Patnala R, Clements J & Batra J (2013). Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genet* 14:39.

Patra SK, Patra A & Dahiya R (2001). Histone deacetylase and DNA methyltransferase in human prostate cancer. *Biochem Biophys Res Commun* 287:705-713.

Peltonen L, Jalanko A & Varilo T (1999). Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* 8:1913-1923.

Petersen B, Petersen TN, Andersen P, et al. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 9:51.

Pickrell JK, Marioni JC, Pai AA, et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768-772.

Pico AR, Kelder T, van Iersel MP, et al. (2008). WikiPathways: pathway editing for the people. *PLoS Biol* 6:e184.

Picollo A, Malvezzi M & Accardi A (2015). TMEM16 proteins: unknown structure and confusing functions. *J Mol Biol* 427:94-105.

Pierce BL, Friedrichsen-Karyadi DM, McIntosh L, et al. (2007). Genomic scan of 12 hereditary prostate cancer families having an occurrence of pancreas cancer. *Prostate* 67:410-415.

Pomerantz MM & Freedman ML (2011). The genetics of cancer risk. *Cancer J* 17:416-422.

Pomerantz M & Freedman ML (2013). Clinical uncertainty of prostate cancer genetic risk panels. *Sci Transl Med* 5:182ed6.

Purcell S, Neale B, Todd-Brown K, et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-575.

Quinlan AR & Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.

Quinonez SC & Innis JW (2014). Human HOX gene disorders. *Mol Genet Metab* 111:4-15.

Ren G, Zhang G, Dong Z, et al. (2009). Recruitment of HDAC4 by transcription factor YY1 represses HOXB13 to affect cell growth in AR-negative prostate cancers. *Int J Biochem Cell B* 41:1094-1101.

Richards S, Aziz N, Bale S, et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17:405-424.

Rizzo JM & Buck MJ (2012). Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prev Res* 5:887-900.

Rubin MA (2015). Toward a prostate cancer precision medicine. *Urol Oncol* 33:73-74.

Rökman A, Ikonen T, Mononen N, et al. (2001). ELAC2/HPC2 involvement in hereditary and sporadic prostate cancer. *Cancer Res* 61:6038-6041.

Rökman A, Baffoe-Bonnie AB, Gillanders E, et al. (2005). Hereditary prostate cancer in Finland: fine-mapping validates 3p26 as a major predisposition locus. *Hum Genet* 116:43-50.

Saaristo L, Wahlfors T, Lilja H, et al. Genetic testing in identification of BPH patients in risk of developing prostate cancer. *Manuscript.*

Sahu SK, Gummadi SN, Manoj N, et al. (2007). Phospholipid scramblases: an overview. *Arch Biochem Biophys* 462:103-114.

Sakoda LC, Jorgenson E & Witte JS (2013). Turning of COGS moves forward findings for hormonally mediated cancers. *Nat Genet* 45:345-348.

Salovaara R, Loukola A, Kristo P, et al. (2000). Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J Clin Oncol* 18:2193-2200.

Sandhu GS & Andriole GL (2012). Overdiagnosis of prostate cancer. *J Natl Cancer Inst Monogr* 2012:146-151.

Schaid DJ (2004). The complex genetic epidemiology of prostate cancer. *Hum Mol Genet* 13:R103-R121.

Schleutker J, Matikainen M, Smith J, et al. (2000). A genetic epidemiological study of hereditary prostate cancer (HPC) in Finland: frequent HPCX linkage in families with late-onset disease. *Clin Cancer Res* 6:4810-4815.

Schleutker J, Baffoe-Bonnie AB, Gillanders E, et al. (2003). Genome-wide scan for linkage in Finnish hereditary prostate cancer (HPC) families identifies novel susceptibility loci at 11q14 and 3p25-26. *Prostate* 57:280-289.

Schröder FH, Hugosson J, Roobol MJ, et al. (2009). Screening and prostate-cancer mortality in a randomized European study. *New Engl J Med* 360:1320-1328.

Schumacher FR, Berndt SI, Siddiq A, et al. (2011). Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet* 20:3867-3875.

Schwartz GG (2014). Vitamin D in blood and risk of prostate cancer: lessons from the selenium and vitamin E cancer prevention trial and the prostate cancer prevention trial. *Cancer Epidemiol Biomarkers Prev* 23:1447-1449.

Schwarz JM, Cooper DN, Schuelke M, et al. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 11:361-362.

Seppälä EH, Ikonen T, Mononen N, et al. (2003a). CHEK2 variants associate with hereditary prostate cancer. *Brit J Cancer* 89:1966-1970.

Seppälä EH, Ikonen T, Autio V, et al. (2003b). Germ-line alterations in MSR1 gene and prostate cancer risk. *Clin Cancer Res* 9:5252-5256.

Sfanos KS & De Marzo AM (2012). Prostate cancer and inflammation: the evidence. *Histopathology* 60:199-215.

Sfanos KS, Isaacs WB & De Marzo AM (2013). Infections and inflammation in prostate cancer. *Am J Clin Exp Urol* 1:3-11.

Shan J, Al-Rumaihi K, Rabah D, et al. (2013). Genome scan study of prostate cancer in Arabs: identification of three genomic regions with multiple prostate cancer susceptibility loci in Tunisians. *J Transl Med* 11:121.

Sharma A, Yeow WS, Ertel A, et al. (2010). The retinoblastoma tumor suppressor controls androgen signaling and human prostate cancer progression. *J Clin Invest* 120:4478-4492.

Shen MM & Abate-Shen C (2010). Molecular genetics of prostate cancer: new prospects for old challenges. *Gene Dev* 24:1967-2000.

Siltanen S, Fischer D, Rantapero T, et al. (2013). ARLTS1 and prostate cancer risk – analysis of expression and regulation. *PLoS ONE* 8:e72040.

Singh AP, Bafna S, Chaudhary K, et al. (2008). Genome-wide expression profiling reveals transcriptomic variation and perturbed gene networks in androgen-dependent and androgen-independent prostate cancer cells. *Cancer Lett* 259:28-38.

Smith SC, Palanisamy N, Zuhlke KA, et al. (2014). HOXB13 G84E-related familial prostate cancers: a clinical, histologic, and molecular survey. *Am J Surg Pathol* 38:615-626.

Spans L, Clinckemalie L, Helsen C, et al. (2013). The genomic landscape of prostate cancer. *Int J Mol Sci* 14:10822-10851.

Stamey TA, Yang N, Hay AR, et al. (1987). Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. *New Engl J Med* 317:909-916.

Stankiewicz P & Lupski JR (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med* 61:437-455.

Stark M & Hayward N (2007). Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. *Cancer Res* 67:2632-2642.

Steinberg GD, Carter BS, Beaty TH, et al. (1990). Family history and the risk of prostate cancer. *Prostate* 17:337-347.

Stelzer G, Dalah I, Stein TI, et al. (2011). In-silico human genomics with GeneCards. *Hum Genomics* 5:709-717.

Stenson PD, Mort M, Ball EV, et al. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133:1-9.

Strand SH, Orntoft TF & Sorensen KD (2014). Prognostic DNA methylation markers for prostate cancer. *Int J Mol Sci* 15:16544-16576.

Stranger BE, Forrest MS, Dunning M, et al. (2007a). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315:848-853.

Stranger BE, Nica AC, Forrest MS, et al. (2007b). Population genomics of human gene expression. *Nat Genet* 39:1217-1224.

Suarez BK, Lin J, Burmester JK, et al. (2000). A genome screen of multiplex sibships with prostate cancer. *Am J Hum Genet* 66:933-944.

Sudmant PH, Rausch T, Gardner EJ, et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75-81.

Sulonen AM, Ellonen P, Almusa H, et al. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 12:R94.

Syrjäkoski K, Vahteristo P, Eerola H, et al. (2000). Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients. *J Natl Cancer Inst* 92:1529-1531.

Szklarczyk D, Franceschini A, Wyder S, et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43:D447-D452.

Tao ZQ, Shi AM, Wang KX, et al. (2015). Epidemiology of prostate cancer: current status. *Eur Rev Med Pharmacol Sci* 19:805-812.

Tavtigian SV, Simard J, Teng DH, et al. (2001). A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 27:172-180.

Teerlink CC, Thibodeau SN, McDonnell SK, et al. (2014). Association analysis of 9,560 prostate cancer cases from the International Consortium of Prostate Cancer Genetics confirms the role of reported prostate-cancer associated SNPs for familial disease. *Hum Genet* 133:347-356.

Teles Alves I, Hartjes T, McClellan E, et al. (2015). Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. *Oncogene* 34:568-577.

Thomas G, Jacobs KB, Yeager M, et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* 40:310-315.

Todd R & Wong DT (1999). Oncogenes. *Anticancer Res* 19:4729-4746.

Tomlins SA, Rhodes DR, Perner S, et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310:644-648.

Trapnell C, Pachter L & Salzberg SL (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25:1105-1111.

Umar A & Kunkel TA (1996). DNA-replication fidelity, mismatch repair and genome instability in cancer cells. *Eur J Biochem* 238:297-307.

Varambally S, Dhanasekaran SM, Zhou M, et al. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 419:624-629.

Venkatachalam R, Verwiel ET, Kamping EJ, et al. (2011). Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int J Cancer* 129:1635-1642.

Veyrieras JB, Kudaravalli S, Kim SY, et al. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 4:e1000214.

Vickers AJ, Sjoberg DD, Ulmert D, et al. (2014). Empirical estimates of prostate cancer overdiagnosis by age and prostate-specific antigen. *BMC Med* 12:26.

Villers A, McNeal JE, Freiha FS, et al. (1992). Multiple cancers in the prostate. Morphologic features of clinically recognized versus incidental tumors. *Cancer* 70:2313-2318.

Visakorpi T, Hyytinen E, Koivisto P, et al. (1995). In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nat Genet* 9:401-406.

Vogelstein B, Papadopoulos N, Velculescu VE, et al. (2013). Cancer genome landscapes. *Science* 339:1546-1558.

Wain HM, Bruford EA, Lovering RC, et al. (2002). Guidelines for human gene nomenclature. *Genomics* 79:464-470.

Wang AH, Bertos NR, Vezmar M, et al. (1999). HDAC4, a human histone deacetylase related to yeast HDA1, is a transcriptional corepressor. *Mol Cell Biol* 19:7816-7827.

Wang K, Li M, Hadley D, et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17:1665-1674.

Wang J, Duncan D, Shi Z, et al. (2013). Web-based gene set analysis toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 41:W77-W83.

Wang Z, Qin G & Zhao TC (2014). Histone deacetylase 4 (HDAC4): mechanism of regulations and biological functions. *Epigenomics* 6:139-150.

Weichenhan D & Plass C (2013). The evolving epigenome. *Hum Mol Genet* 22:R1-R6.

Welter D, MacArthur J, Morales J, et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42:D1001-D1006.

Williams JL, Greer PA & Squire JA (2014). Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data. *Cancer Genet* 207:474-488.

Witte JS (2009). Prostate cancer genomics: towards a new understanding. *Nat Rev Genet* 10:77-82.

Woenckhaus J & Fenic I (2008). Proliferative inflammatory atrophy: a background lesion of prostate cancer? *Andrologia* 40:134-137.

Wolters T, Montironi R, Mazzucchelli R, et al. (2012). Comparison of incidentally detected prostate cancer with screen-detected prostate cancer treated by prostatectomy. *Prostate* 72:108-115.

Wright FA, Sullivan PF, Brooks AI, et al. (2014). Heritability and genomics of gene expression in peripheral blood. *Nat Genet* 46:430-437.

Wu R, Wang H, Wang J, et al. (2014a). EphA3, induced by PC-1/PrLZ, contributes to the malignant progression of prostate cancer. *Oncol Rep* 32:2657-2665.

Wu C, Zhu C & Jegga AG (2014b). Integrative literature and data mining to rank disease candidate genes. *Methods Mol Biol* 1159:207-226.

Xi HQ & Zhao P (2011). Clinicopathological significance and prognostic value of EphA3 and CD133 expression in colorectal carcinoma. *J Clin Pathol* 64:498-503.

Xu J, Meyers D, Freije D, et al. (1998). Evidence for a prostate cancer susceptibility locus on the X chromosome. *Nat Genet* 20:175-179.

Xu J, Zheng SL, Komiya A, et al. (2002). Germline mutations and sequence variants of the macrophage scavenger receptor 1 gene are associated with prostate cancer risk. *Nat Genet* 32:321-325.

Xu J, Dimitrov L, Chang BL, et al. (2005). A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the International Consortium for Prostate Cancer Genetics. *Am J Hum Genet* 77:219-229.

Xu J, Lange EM, Lu L, et al. (2013). HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). *Hum Genet* 132:5-14.

Xu X, Hussain WM, Vijai J, et al. (2014). Variants at IRX4 as prostate cancer expression quantitative trait loci. *Eur J Hum Genet* 22:558-563.

Yang Y, Tse AKW, Li P, et al. (2011). Inhibition of androgen receptor activity by histone deacetylase 4 through receptor SUMOylation. *Oncogene* 30:2207-2218.

Yeager M, Orr N, Hayes RB, et al. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 39:645-649.

Zeegers MP, Jellema A & Ostrer H (2003). Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma: a meta-analysis. *Cancer* 97:1894-1903.

Zeller T, Wild P, Szymczak S, et al. (2010). Genetics and beyond – the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* 5:e10693.

Zheng SL, Sun J, Wiklund F, et al. (2008). Cumulative association of five genetic variants with prostate cancer. *New Engl J Med* 358:910-919.

Zhou Y, Bolton EC & Jones JO (2015). Androgens and androgen receptor signaling in prostate tumorigenesis. *J Mol Endocrinol* 54:R15-R29.

Zhu C, Wu C, Aronow BJ, et al. (2014). Computational approaches for human disease gene prediction and ranking. *Adv Exp Med Biol* 799:69-84.

Zuhlke KA, Johnson AM, Tomlins SA, et al. (2014). Identification of a novel germline SPOP mutation in a family with hereditary prostate cancer. *Prostate* 74:983-990.

Özdemir BC, Hensel J, Secondini C, et al. (2014). The molecular signature of the stroma response in prostate cancer-induced osteoblastic bone metastasis highlights expansion of hematopoietic and prostate epithelial stem cell niches. *PLoS ONE* 9:e114530.

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

Cancer Epidemiology, Biomarkers & Prevention

Research Article

HOXB13 G84E Mutation in Finland: Population-Based Analysis of Prostate, Breast, and Colorectal Cancer Risk

Virpi H. Laitinen[1], Tiina Wahlfors[1], Leena Saaristo[1], Tommi Rantapero[1], Liisa M. Pelttari[6], Outi Kilpivaara[7], Satu-Leena Laasanen[2,3], Anne Kallioniemi[1], Heli Nevanlinna[6], Lauri Aaltonen[7], Robert L. Vessella[9], Anssi Auvinen[4], Tapio Visakorpi[1], Teuvo L.J. Tammela[5], and Johanna Schleutker[1,8]

Abstract

Background: A recently identified germline mutation G84E in *HOXB13* was shown to increase the risk of prostate cancer. In a family-based analysis by The International Consortium for Prostate Cancer Genetics (ICPCG), the G84E mutation was most prevalent in families from the Nordic countries of Finland (22.4%) and Sweden (8.2%).

Methods: To further investigate the importance of G84E in the Finns, we determined its frequency in more than 4,000 prostate cancer cases and 5,000 controls. In addition, 986 breast cancer and 442 colorectal cancer (CRC) cases were studied. Genotyping was conducted using TaqMan, MassARRAY iPLEX, and sequencing. Statistical analyses were conducted using Fisher exact test, and overall survival was analyzed using Cox modeling.

Results: The frequency of the G84E mutation was significantly higher among patients with prostate cancer and highest among patients with a family history of the disease, hereditary prostate cancer [8.4% vs. 1.0% in controls; OR 8.8; 95% confidence interval (CI), 4.9–15.7]. The mutation contributed significantly to younger age ($\leq$55 years) at onset and high prostate-specific antigen (PSA; $\geq$20 ng/mL) at diagnosis. An association with increased prostate cancer risk in patients with prior benign prostate hyperplasia (BPH) diagnosis was also revealed. No statistically significant evidence for a contribution in CRC risk was detected, but a suggestive role for the mutation was observed in familial BRCA1/2-negative breast cancer.

Conclusions: These findings confirm an increased cancer risk associated with the G84E mutation in the Finnish population, particularly for early-onset prostate cancer and cases with substantially elevated PSA.

Impact: This study confirms the overall importance of the *HOXB13* G84E mutation in prostate cancer susceptibility. *Cancer Epidemiol Biomarkers Prev; 22(3); 452–60. ©2012 AACR.*

Introduction

In 2010, more than 4,700 Finnish men were diagnosed with prostate cancer and 847 died of it. These figures make the disease the most commonly diagnosed cancer in Finland and the second most common cause of cancer-related death (1). Despite its high incidence and mortality rates, the exact molecular mechanisms underlying the initiation and progression of prostate cancer still remain largely unknown.

Worldwide, compelling evidence has accumulated in favor of a significant but heterogeneous genetic component in prostate cancer susceptibility. On the basis of twin studies, heritability has been estimated as high as 16% to 45% (2, 3). However, the genetics of prostate cancer has proven hard to dissect. So far, only a few risk genes have been identified, although approximately 40 loci have been associated to genetic susceptibility (4, 5). Rare Mendelian genes with high penetrance, such as ribonuclease L [*RNASEL* (MIM 180435); ref. 6], explain perhaps 5% of prostate cancer susceptibility, whereas the more common genetic variants found in genome-wide association studies (GWAS) explain only approximately 25% of familial risk (7). Although GWAS have discovered many loci associated with prostate cancer risk, single-nucleotide polymorphisms (SNP) related to clinical outcome, that is, disease aggressiveness, have not been found. Consequently, there is renewed interest in family studies

Authors' Affiliations: [1]Institute of Biomedical Technology/BioMediTech, University of Tampere and Fimlab Laboratories, Tampere, Finland; [2]Department of Pediatrics, Genetics Outpatient Clinic, Tampere University Hospital, Tampere, Finland; [3]Department of Dermatology, Tampere University Hospital, Tampere, Finland; [4]Department of Epidemiology, School of Health Sciences, University of Tampere, Tampere, Finland; [5]Department of Urology, Tampere University Hospital and Medical School, University of Tampere, Tampere, Finland; [6]Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland; [7]Department of Medical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland; [8]Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, Turku, Finland; and [9]Department of Urology, University of Washington Medical Center, Seattle, Washington, USA

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

V.H. Laitinen and T. Wahlfors contributed equally to this work.

Corresponding Author: Johanna Schleutker, Medical Biochemistry and Genetics, Institute of Biomedicine, Kiinamyllynkatu 10, FI-20014 University of Turku, Finland. Phone: 358-2-3337453; Fax: 358-2-2301280; E-mail: Johanna.Schleutker@utu.fi

doi: 10.1158/1055-9965.EPI-12-1000-T

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

because of the type of information they offer, especially when trying to isolate rare high-impact variants.

Linkage analyses of hereditary prostate cancer (HPC) families have detected a significant signal at the chromosomal region of 17q21-22 in both North American and Finnish populations (8–10). Recently, Ewing and colleagues (11) used targeted next generation sequencing of this region to identify a rare but recurrent germline missense mutation c.251G→A (p.G84E, rs138213197) in the first exon of the homeobox B13 [*HOXB13* (MIM 604607)] gene. This mutation was associated with a significantly increased risk of early-onset, familial prostate cancer.

The *HOXB13* gene belongs to a group of highly conserved homeobox genes that are essential for vertebrate embryogenesis. In humans, there are 4 *HOX* gene clusters (A–D) in separate chromosomes, and the *HOXB* cluster is localized in the 17q21-22 region (12). *HOXB13* is highly expressed in both normal and cancerous prostate. The HOXB13 protein is a sequence-specific, 284-amino acid transcription factor that interacts with androgen receptor and has an important role in prostate development (13). It has been shown to regulate cellular responses to androgens, such as promotion of androgen-independent growth in prostate cancer cell lines (14) by activating or repressing the expression of most androgen receptor–responsive genes (15). In addition to prostate cancer, HOXB13 has also been shown to have a role as a tumor suppressor in primary colorectal cancers (CRC; ref. 16), and it predicts breast cancer recurrence (17) and tamoxifen response (18).

Given the linkage evidence to the 17q21-22 locus in Finnish prostate cancer families (10), and the exceptionally high proportion of Finnish families with the G84E mutation, as shown in a recent International Consortium for Prostate Cancer Genetics (ICPCG) study (19), we genotyped the G84E mutation in 4,571 prostate cancer cases and 5,467 controls, together with 516 benign prostate hyperplasia samples, 10 prostatic cell lines, and 19 LuCaP xenografts. We also investigated its role in prostate cancer risk, clinical outcome, and survival. To evaluate the cancer specificity of G84E in the genetically homogeneous Finnish population, we analyzed an additional 3,336 samples collected from breast and CRC cases and controls.

Materials and Methods

Study subjects

All cancer cases and controls genotyped in this study were of Finnish origin. Written informed consent was obtained from each study subject. The cancer diagnosis was confirmed from medical records. The study protocol was approved by the research ethics committee at Pirkanmaa Hospital District (Tampere, Finland) and by the National Supervisory Authority for Welfare and Health. Different sample types included in the analyses are presented in the Supplementary Table S1.

Prostate cancer. A total of 4,571 Finnish prostate cancer samples were genotyped. Of these, 3,197 unselected cases were collected in the Pirkanmaa Hospital District. Another unselected set of subjects consisted of 1,184 Finnish cancer cases recruited by the Finnish arm of The European Randomized Study of Screening for Prostate Cancer. This study was initiated in the early 1990s to evaluate the effect of prostate-specific antigen (PSA) screening on death rates from prostate cancer (20). In addition to the unselected cases, genotype data for 190 index cases derived from Finnish prostate cancer families were included. The collection of the Finnish familial prostate cancer families has been described previously (21, 22). All of the 190 families used in this study had at least 2 members affected by prostate cancer, with the majority of families ($n = 151$) having at least 3 confirmed cases. All affected persons were either first- or second-degree relatives of the index cases. Only an index case was originally genotyped, and additional individuals were studied only to confirm segregation of the mutation. Seventy-six index individuals overlapped with those genotyped in the large multinational ICPCG study (19). To investigate the cosegregation of the G84E mutation in nonoverlapping, mutation-positive families, additional healthy and affected family members were genotyped. The most representative clinical features for each of the 3 prostate cancer patient groups are summarized in Table 1.

Germline DNA was also available from 516 clinically and pathologically defined cases of benign prostate hyperplasia from the Urology Outpatient Clinic in Tampere University Hospital (Tampere, Finland; BPH; samples collected in 1998–2004): 254 of these cases were later diagnosed with prostate cancer. In addition to germline DNA samples, the G84E status was analyzed in 2 normal cell lines (PrEC and EP156T), 8 prostate cancer cell lines (LAPC4, LNCaP, DuCaP, DU145, PC-3, VcaP, and 2 separate lines, 22Rv1 and CWR22Pc, derived from CWR22), and 19 LuCaP xenografts. DU145, PC-3, 22Rv1, and LNCaP were obtained from the American Type Culture Collection. CWR22Pc was provided by Dr. Marja Nevalainen (Thomas Jefferson University, Philadelphia, PA). LAPC4 was obtained from Dr. Charles Sawyers (University of California at Los Angeles, Los Angeles, CA). VCaP and DuCaP were obtained from Dr. Jack Schalken (Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands). PrEC was obtained from Lonza (Lonza Walkersville). EP156T was kindly provided by Dr. Varda Rotter (Weizmann Institute of Science, Rehovot, Israel).

Breast cancer. Tampere subgroup: 86 index cases from well-characterized high-risk breast cancer families were genotyped. In these families, patients with breast cancer were diagnosed at an early age or at least 3 first-degree relatives had breast or ovarian cancer. The sample set is described in more detail elsewhere (23). In addition, 410 unselected Finnish breast cancer cases, described previously by Syrjäkoski and colleagues

Table 1. Clinical characteristics of the 3 prostate cancer patient groups analyzed in this study

Characteristics	Variables	All FAM%[a] (n)	UNS%[b] (n)	SCRcase%[c] (n)
Average age at onset	Age at onset (y)	62.8	68.6	67.0
Prostate specific antigen	≤4.0 ng/mL	5.4 (9)	8.0 (234)	12.9 (152)
	4.1–9.9 ng/mL	35.5 (59)	43.0 (1,258)	61.1 (719)
	10.0–19.9 ng/mL	26.5 (44)	25.3 (740)	17.9 (211)
	20.0–49.9 ng/mL	21.1 (35)	13.3 (389)	6.5 (77)
	50.0–99.9 ng/mL	4.8 (8)	4.7 (137)	0.8 (9)
	≥100 ng/mL	6.6 (11)	5.7 (167)	0.8 (9)
	Missing data	12.4 (24)	8.5 (272)	0.6 (7)
Primary treatment	Prostatectomy	46.0 (82)	34.7 (1,030)	23.0 (32)
	Radiotherapy	16.9 (30)	18.4 (546)	39.1 (55)
	Hormonal therapy	30.9 (55)	37.9 (1,124)	9.8 (14)
	Active surveillance	4.5 (8)	5.6 (166)	14.7 (21)
	Brachytherapy	1.7 (3)	2.9 (86)	12.6 (18)
	Cystectomy	—	0.5 (15)	0.7 (1)
	Missing data	6.3 (12)	7.2 (230)	88.0 (1,043)
Gleason score for biopsy	3	2.7 (4)	2.7 (72)	2.5 (29)
	4	11.6 (17)	4.1 (109)	8.8 (102)
	5	15.7 (23)	11.4 (304)	12.3 (143)
	6	32.7 (48)	36.5 (972)	42.8 (496)
	7	22.4 (33)	27.8 (740)	25.2 (292)
	8	8.2 (12)	8.4 (224)	6.0 (70)
	9	6.0 (9)	8.4 (224)	2.0 (23)
	10	0.7 (1)	0.7 (19)	0.4 (5)
	Missing data	22.6 (43)	16.7 (534)	2.1 (25)
Progression	PSA progression	13.7 (26)	30.9 (988)	—
Cause of death	Overall deaths	42.1 (80)	43.1 (1,378)	8.8 (104)
	Prostate cancer	35.0 (67)	26.6 (850)	5.7 (67)

[a] All FAM, familial index cases from all 190 Finnish prostate cancer families.

[b] UNS, unselected cases.

[c] SCRcase, screening trial cases.

(24), were analyzed in this study. Helsinki subgroup: genotyping was conducted for 237 familial and 253 patients with sporadic breast cancer. The patients with familial breast cancer were collected at the Helsinki University Central Hospital Departments of Oncology and Clinical Genetics (Helsinki, Finland) as previously described (25). They had a strong familial background of breast cancer with 3 or more breast or ovarian cancers among first- or second-degree relatives, including the proband. The patients with sporadic breast cancer were part of an unselected series collected at the Helsinki University Central Hospital Department of Surgery in 2001 to 2004 (26). In both the Tampere and Helsinki subgroups, all of the patients with familial breast cancer tested negative for *BRCA1* (MIM 113705) and *BRCA2* (MIM 600185) founder mutations.

Colorectal cancer. The sample set consisted of 442 CRC cases belonging to a Finnish population-based series of 1,042 patients with CRC. Fifty-seven CRC cases were classified as familial, having at least 1 first-degree relative with CRC. The data were collected prospectively at 9 Finnish central hospitals between 1994 and 1998 as described by Aaltonen and colleagues (27) and Salovaara and colleagues (28).

Controls. All control subjects for breast cancer and CRC, as well as the population control group for prostate cancer, consisted of population-matched healthy individuals of ages between 18 and 65 years. The blood DNA samples were obtained from the Finnish Red Cross Blood Transfusion Service. Population control subjects for prostate cancer included 923 anonymous male blood donors. Breast cancer controls for the Tampere and Helsinki subgroups comprised 900 and 549 anonymous, healthy female blood donors, respectively. Blood-derived DNA samples from an additional 459 healthy individuals were used as CRC controls.

Prostate cancer control subjects ($n = 4{,}544$) belonging to the screening trial control group were derived from the Finnish arm of the European Randomized Study of Screening for Prostate Cancer (20). All members of this control group were age-standardized (from 59 to 79 years) healthy men who had undergone PSA screening. The

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

disease status is annually evaluated from the records of the Finnish Cancer Registry.

SNP genotyping

Prostate and breast cancer samples, as well as the cell lines and xenografts, were genotyped for the G84E mutation (rs138213197) using a Custom TaqMan SNP assay (Applied Biosystems/Life Technologies) according to the manufacturer's instructions. Duplicate test samples and 4 negative controls were included in each 384-well plate.

BPH samples were genotyped by the Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki (Helsinki, Finland) using the MassARRAY iPLEX platform (Sequenom, Inc.).

DNA sequencing

The mutation was confirmed in a selected set of prostate and breast cancer samples by standard Sanger sequencing using an ABI PRISM BigDye Termination Cycle Sequencing Ready Reaction Kit (Applied Biosystems/Life Technologies). CRC cases and controls were genotyped by sequencing the coding exons of *HOXB13*. CRC DNA from all 7 G84E carriers was extracted from freshly frozen tissue, and the coding region of *HOXB13* was sequenced for LOH analysis. Primer sequences are available upon request.

Statistical analysis

The statistical significance of the association between the *HOXB13* G84E mutation and prostate cancer, breast cancer, or CRC was evaluated using a Fisher exact test, implemented in PLINK (29) and GraphPad Prism 5.02 (GraphPad Software, Inc.) softwares. In addition to case–control comparisons, case–case analyses evaluated the impact of the mutation to the clinical features (PSA, Gleason score, age at onset, and progression). All *P* values were 2-sided. The association between the mutation and overall survival was analyzed using a Cox model. Survival time (years) after diagnosis was compared between carriers and noncarriers. Statistical significance of the survival differences between the G84E carriers and noncarriers were calculated with log-rank and Gehan–Breslow–Wilcoxon tests.

In silico pathogenicity prediction

The pathogenicity of G84E was evaluated by using a machine learning-based method PON-P (Pathogenic-or-Not Pipeline; ref. 30) that includes 6 independent tolerance predictors (SIFT, PolyPhen-2, SNAP, PHD-SNP, PANTHER, and SNP&GO) and the pipeline's own metapredictor, which integrates the output of 5 predictors (SIFT, SNAP, PolyPhen-2, PHD-SNP, and I-Mutant-3) as the input to make the pathogenicity prediction. Two additional programs, NetSurfP (31) and SABLE 2 (32), were used to investigate the sequence environment of G84. These programs predict features such as the secondary structure, transmembrane regions, and the relative solvent accessibilities of the amino acids based on the amino acid sequence of the given protein. Protein stability was examined using the I-Mutant-3 (33) and MuPro (34) programs, also implemented in PON-P, and an additional program called iPTREE-STAB (35).

Results

Prostate cancer

The overall call rate of the mutation site among prostate cancer samples was 99.8%, and the average concordance of duplicated samples was 99.9%. The G84E mutation was in Hardy–Weinberg equilibrium in both cases and controls. The overall minor allele frequency in the entire sample set was 1.9%. The G84E mutation was detected in 188 subjects, of which 160 were patients with prostate cancer (carrier frequency 3.5%) and 28 were healthy controls (0.5%). Of the cases carrying G84E, 3.4% (155 of 4,571) were heterozygous, and 0.1% (5 of 4,571) were homozygous for the mutation. The observed G84E carrier frequency for the unselected cases from the Pirkanmaa Hospital District was 3.6% (114 of 3,197), but the frequency was only 2.2% (26 of 1,184) for the screening trial patients. The highest carrier frequency of 8.4% (16 of 190) was observed among index patients with a positive family history of prostate cancer. In this group, the case subjects were significantly more likely to carry the mutation compared with population controls [carrier frequency 1.0%; $P = 2.318$ e-18; OR, 8.8; 95% confidence interval (CI), 4.9–15.7]. In addition, statistically significantly higher carrier frequencies were detected among cases with a positive family history of prostate cancer compared with unselected cases ($P = 1.982$e-06; OR, 2.5; 95% CI, 1.7–3.6). Table 2 summarizes the results of the association analyses.

Case–case analysis of the G84E mutation in relation to clinical features of prostate cancer revealed a significant association with younger age (≤55 years) at diagnosis ($P = 0.0008$; OR, 2.0; 95% CI, 1.3–3.0). Likewise, carrier frequency was significantly higher among men with serum PSA concentrations 20 ng/mL or more at diagnosis ($P = 0.006$; OR, 1.4; 95% CI, 1.1–1.9). However, no evidence for an association with tumor grade (Gleason score ≥8 vs. ≤6) or prostate cancer progression based on elevated PSA (present vs. absent) was observed (Table 3). Gleason 7 was left out of the analysis to decrease the heterogeneity of the compared groups because it was not possible to differentiate Gleason scores of 7 as either "3+4" or "4+3." A slightly but not significantly poorer overall survival (HR, 1.16; 95% CI, 0.9–1.5) was observed in mutation carriers relative to noncarriers. A significantly elevated risk of prostate cancer was found to be associated with the G84E mutation in a group of patients with prior BPH diagnosis ($P = 0.01084$; OR, 4.6; 95% CI, 1.3–16.2). Interestingly, none of the prostate cell lines or LuCaP xenografts carried the A allele of the mutation.

Of the 190 Finnish prostate cancer families included in this study, 32 indexes (17%) were found to be carriers of the G84E mutation. Fifteen of these 32 families

Table 2. Summary of results obtained from the case–control and case–case association analyses of the G84E mutation and prostate cancer risk

Prostate cancer datasets	F_A%	F_U%	*P* value	OR (95% CI)
All cases and controls	3.5	0.5	1.1×10^{-62}	7.1 (5.5–9.3)
UNS[a] vs. Pco[b]	3.6	1.0	1.8×10^{-8}	3.6 (2.2–5.7)
UNS vs. SCRco[c]	3.6	0.3	6.2×10^{-57}	13.4 (8.9–20.3)
SCRcase[d] vs. SCRco	2.2	0.3	1.1×10^{-23}	8.0 (4.9–12.9)
SCRcase vs. Pco	2.2	1.0	0.004603	2.1 (1.2–3.6)
All FAM[e] vs. Pco	8.4	1.0	2.3×10^{-18}	8.8 (4.9–15.7)
All FAM vs. SCRco	8.4	0.3	1.8×10^{-89}	33.1 (19.4–56.5)
All FAM vs. UNS	8.4	3.6	2.0×10^{-6}	2.5 (1.7–3.6)
All FAM vs. SCRcase	8.4	2.2	4.2×10^{-11}	4.2 (2.6–6.6)
FAM[f] vs. Pco	7.9	1.0	1.5×10^{-13}	8.2 (4.3–16.0)
FAM vs. SCRco	7.9	0.3	4.4×10^{-63}	31.1 (16.7–57.8)
FAM vs. UNS	7.9	3.6	0.0006835	2.3 (1.4–3.8)
FAM vs. SCRcase	7.9	2.2	2.6×10^{-7}	3.9 (2.2–6.8)
BPHcase[g] vs. BPHco[h]	2.6	0.6	0.011	4.6 (1.3–16.2)

NOTE: F_A and F_U represent the frequencies of G84E carriers among affected and unaffected subjects, respectively. All *P* values are statistically significant.

[a] UNS, unselected cases.

[b] Pco, population controls.

[c] SCRco, screening trial controls.

[d] SCRcase, screening trial cases.

[e] All FAM, familial index cases from all 190 Finnish prostate cancer families.

[f] FAM, familial index cases from the 114 Finnish prostate cancer families analyzed in this study (the 76 familial cases overlapping with the ICPCG dataset are omitted).

[g] BPHcase, patients with BPH with a later diagnosis of prostate cancer.

[h] BPHco, patients with BPH with no diagnosis of prostate cancer.

overlapped with the ICPCG dataset (19). Cosegregation of G84E with prostate cancer in the remaining 17 families was assessed by genotyping an additional 28 healthy and 37 affected family members, for whom DNA samples were available. In 11 of 17 families, the G84E mutation cosegregated with the disease in 20 genotyped cases, representing 53% of the total cancer cases in these families. Segregation of the mutation with the disease was incomplete in 6 families, as both unaffected mutation carriers ($n = 5$) and mutation-negative

Table 3. Summary of results obtained from the case–case association analysis of the G84E mutation and selected clinical features

Age at diagnosis	G84E carriers% (*n*)	G84E noncarriers% (*n*)	*P* value	OR (95% CI)
≤55 y	6.25 (13)	93.75 (195)	*0.0007959*	2.0 (1.3–3.0)
>55 y	3.40 (148)	96.60 (4,209)		
PSA at diagnosis				
≥20 ng/mL	4.56 (39)	95.44 (816)	*0.006187*	1.4 (1.1–1.9)
<20 ng/mL	3.19 (110)	96.81 (3,336)		
PSA progression				
Present	3.76 (39)	96.24 (997)	0.5034	1.1 (0.8–1.4)
Absent	3.51 (124)	96.49 (3,406)		
Gleason score				
≥8	4.04 (22)	95.96 (523)	0.09918	1.3 (1.0–1.9)
≤6	3.11 (70)	`96.89 (2,182)		

NOTE: The statistically significant *P* values are italicized.

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

patients ($n = 7$) were observed. (segregation presented in the Supplementary Table S2).

Breast cancer

The G84E mutation was identified in 6 of 323 (1.9%) of the familial breast cancer cases, 10 of 663 (1.5%) of the sporadic breast cancer cases and 16 of 1,449 (1.1%) of the population controls. Case–control association analyses were conducted for the entire dataset and separately for both subgroups (familial and sporadic), but no statistically significant differences in carrier frequencies between cases and controls were observed (data not shown). However, in the high-risk, familial Tampere subgroup, the frequency of G84E carriers was 3.5%, a figure similar to the number of mutation carriers among the Finnish patients with prostate cancer. The OR of 3.2 (95% CI, 0.9–11.9) is suggestive of an association between G84E and increased breast cancer risk.

Colorectal cancer

Of the 442 patients with CRC, 7 (1.6%) were identified as carriers of the *HOXB13* G84E mutation, and none of these were familial. No evidence of allelic imbalance was observed in the LOH analysis of the G84E-positive tumors. In a case–control association analysis, the difference in carrier frequencies was nonsignificant between cases (1.6%) and population controls (0.9%).

In silico analysis

To further explore the mechanistic function of G84E in prostate cancer risk, *in silico* analyses were conducted. In a pathogenicity prediction analysis, the G84E mutation was predicted to be deleterious in 5 of 6 of the tolerance predictors included in PON-P. However, the pipeline's own meta-predictor indicated the mutation to be tolerated. NetSurfP and SABLE 2 estimated glysine 84 to be located in a region buried inside the protein structure. Moreover, the sequence surrounding glysine 84 was found to be relatively hydrophobic, suggesting that G84 is located in the hydrophobic core of HOXB13. However, the applied programs gave conflicting results in protein stability tests. (all results presented in the Supplementary Tables S3–S5 and Supplementary Fig. S1).

Discussion

The present results validate an important role for *HOXB13* G84E in prostate cancer predisposition. In the Finnish population, the mutation was detected in 3.5% of all cases and 8.4% of familial prostate cancer, which suggests that G84E may be the strongest genetic marker of prostate cancer reported to date. In the original article by Ewing and colleagues (11), the highest *HOXB13* G84E carrier frequency of 3.1% was observed among men with a positive family history of prostate cancer and an early age (≤55 years) at diagnosis. In Finland, the carrier frequency of the mutation among familial prostate cancer cases was almost 3-fold higher (8.4%), a frequency that is strikingly similar to the carrier frequency observed in Swedish prostate cancer families (8.2%). Moreover, both the Finnish and Swedish mutation carrier families share a common rare haplotype indicating a likely founder effect for the mutation (19). Founder mutations are typical for isolated populations, such as the Finnish population, and they may explain a major fraction of all mutations in specific genes (36, 37). In the Finnish population, strong founder mutations have been detected in breast cancers and CRCs (38, 39). In Finland, founder mutations are often present in geographic clusters when the birthplaces of ancestors are known (40). Here, however, the birthplaces of the grandparents of the G84E-positive patients did not show such a pattern, which may indicate a very old origin of the mutation.

Ewing and colleagues (11) reported control subject carrier frequencies to vary between 0.1% and 0.2%. In our study, the frequency distribution of the G84E mutation in different prostate cancer control groups ranged from 0.3% to 1.0%. The lowest frequency (0.3%) was detected in the age-matched, PSA-screened control group. The variations in carrier frequencies are explained by differences in age distributions between control groups, with the oldest subjects belonging to screening trial controls and the youngest to population controls.

When compared with patients with prostate cancer, G84E carrier frequencies were substantially lower in patients with breast cancer and CRC. On the basis of these results, the *HOXB13* G84E mutation seems to be prostate cancer–specific, although this needs to be verified in larger breast and CRC datasets. However, it is noteworthy that all patients with G84E-positive familial breast cancer tested negative for the Finnish *BRCA1/2* founder mutations, and the highest carrier frequency of 3.5% occurred within the high-risk Tampere subgroup among all studied patients with breast cancer.

Previously, similar prostate cancer–associated mutation carrier frequencies in the Finnish population have been obtained only for mutations 1100delC (3.3%) and I157T (10.8%) in the checkpoint kinase 2 [*CHEK2* (MIM 604373)] gene (41). Analogous to the current *HOXB13* mutation, *CHEK2* mutation frequencies in prostate cancer were significantly higher in populations from Northern and Eastern European countries as compared with North American populations, reflecting population-specific differences (42–44). *CHEK2* is also a known breast cancer risk gene, and the frequency of 1100delC among patients with breast cancer varies similarly between European and North American populations (45, 46).

In previous linkage studies of the 69 Finnish HPC families, a strong signal was observed for the 17q21-22 region (10), and before the present study, G84E was thought to explain this finding. However, of the 32 G84E-positive families analyzed in this study, only 2 families showed linkage (LOD score >0.6) to chromosome 17, suggesting that the G84E-positive and linkage-contributing families are not overlapping. Moreover,

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

cosegregation with prostate cancer was not complete in many of the G84E-positive families, and incomplete penetrance and genetic heterogeneity were observed in 35.3% (6 of 17) of the families, which is consistent with the results of the ICPCG study (19). Of the 5 unaffected mutation carriers observed in this study, 3 were in their sixties and are therefore still at risk for the disease, but the 2 oldest carriers were already 80 and 87 years of age. Contrary to the results reported by Ewing and colleagues (11), we found 5 of the analyzed patients with prostate cancer to be homozygous for the rs138213197 A allele. Two of them represented familial prostate cancer (1 initially reported in the above-mentioned ICPCG study), whereas the other 3 were unselected cases. The 5 homozygous patients did not share any distinctive clinical features relating to disease aggressiveness.

Although G84E seems to explain a considerable fraction of Finnish familial prostate cancer, the linkage signal cannot be explained by *HOXB13* alone and there must be other, yet unidentified genes and variants on chromosome 17 that are responsible for the remaining and quite substantial proportion of HPC cases in Finland. Because of the observed heterogeneity, we evaluated other cancers in the G84E-positive families. In these 32 families, 35 individuals were diagnosed with a cancer other than that of the prostate. Altogether, 17 different cancer types were detected in the patients (10 males and 25 females). No particular cancer type was over-represented. Another cancer was diagnosed in 5 of the patients with G84E-positive prostate cancer, and 5 females had a diagnosis of breast cancer.

Several studies have shown an increased risk of prostate cancer incidence among patients with BPH, although BPH is not considered a premalignant lesion (47, 48). Our collection of BPH cases, from years 1998 to 2004, has been followed-up since and almost half of these cases have been diagnosed with prostate cancer during this follow-up time. In this study, the aim was to assess whether the *HOXB13* G84E mutation has a risk-associated role in prostate cancer occurrence in the BPH cohort. As shown, patients with BPH carrying the G84E mutation were at a significantly increased risk of developing prostate cancer as compared with noncarriers. Because all of these BPH cases were histologically confirmed, there is no chance for misclassification of clinical BPH. Furthermore, the relatively long follow-up time of 8 to 14 years enhances the reliability of the data. Histologic BPH is observed in 50% of men of ages 51 to 60 years and in 70% of men of ages 61 to 70 years (49). Genetic markers that can separate the patients with high-risk BPH from the considerably larger low-risk group would be desirable. Therefore, at least in Finland, G84E deserves serious attention, and genetic testing could be an option for patients with histologically confirmed BPH.

Although numerous genetic variants have been associated with prostate cancer predisposition, their roles as prognostic factors have been limited. Here, the G84E mutation was found to be associated with a high ($\geq$20 ng/mL) PSA concentration at the time of diagnosis, providing evidence for the clinical relevance of G84E in the Finnish population. To our knowledge, this is the first time that G84E has been significantly associated with a clinical feature commonly considered a marker of aggressive disease. However, no difference in other clinical features related to disease aggressiveness, such as Gleason score or prostate cancer progression, was observed between mutation carriers and noncarriers. We also analyzed the association of G84E with overall survival, but the median survival period after prostate cancer diagnosis did not differ between carriers and noncarriers (data not shown). The association of G84E with PSA concentrations may perhaps be explained by a possible regulatory role of HOXB13 on androgen-responsive genes, which warrants further study.

Ewing and colleagues (11) analyzed tumor tissues obtained from G84E carriers and showed that these tumors maintain the expression of *HOXB13*, a finding consistent with the hypothesis that *HOXB13* functions as an oncogene. We confirmed the observation of *HOXB13* expression by analyzing tumor tissue from G84E carriers and noncarriers with immunohistochemistry (data not shown). The pathogenic role of the G84E mutation has not yet been shown by functional studies. We investigated the pathogenicity of G84E using diverse *in silico* predictors. On the basis of our results, it is possible that G84E affects protein stability because a small hydrophobic glycine is replaced with hydrophilic glutamate. To confirm the functionality, *in vivo* studies are needed.

In summary, the rare *HOXB13* mutation has been shown to contribute to prostate cancer risk in Finland, confirming the high frequency of the G84E mutation in this Nordic population. The risk was highest in familial prostate cancer cases. No such effect was observed for CRC, but a suggestive risk effect was detected in a subset of familial breast cancer cases. These results indicate that the G84E mutation may have clinical implications for prostate cancer management in the Finnish population.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: V.H. Laitinen, T. Wahlfors, T.L.J. Tammela, J. Schleutker

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): L. Saaristo, L.M. Pelttari, O. Kilpivaara, S.-L. Laasanen, A. Kallioniemi, H. Nevanlinna, L. Aaltonen, R.L. Vessella, A. Auvinen, T. Visakorpi, T.L.J. Tammela, J. Schleutker

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): V.H. Laitinen, T. Wahlfors, L. Saaristo, T. Rantapero, O. Kilpivaara, H. Nevanlinna, A. Auvinen, T. Visakorpi, T.L.J. Tammela, J. Schleutker

Writing, review, and/or revision of the manuscript: V.H. Laitinen, T. Wahlfors, L. Saaristo, T. Rantapero, L.M. Pelttari, A. Kallioniemi, H. Nevanlinna, L. Aaltonen, R.L. Vessella, A. Auvinen, T. Visakorpi, T.L.J. Tammela, J. Schleutker

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): T. Wahlfors, L. Saaristo, L. Aaltonen, T.L.J. Tammela
Study supervision: T. Wahlfors, T.L.J. Tammela, J. Schleutker

Acknowledgments

The authors thank the patients and families who participated in this study. Riitta Vaalavuo and Riina Liikala are thanked for assistance. The authors also thank Kirsi Kuusisto for her contribution to the collection of the Tampere familial breast cancer patient samples; Drs. Kristiina Aittomäki, Carl Blomqvist, and Karl von Smitten, as well as Sara Vilske, for their help with the Helsinki breast cancer patient samples and data; and Mairi Kuiris and Sini Karjalainen for technical assistance in sequencing the CRC samples.

Grant Support

This work was supported by the Sigrid Juselius Foundation, the Finnish Cancer Organizations, and the Competitive Research Funding of the Pirkanmaa Hospital District (9M094 and 9N069) grants to J. Schleutker; the Academy of Finland grants 116437 and 251074 to J. Schleutker, grant 132473 to H. Nevanlinna, grant 250345 Finnish Center of Excellence in Cancer Genetics Research to L. Aaltonen; and the Helsinki University Hospital Research Funds to H. Nevanlinna.

Received August 29, 2012; revised November 14, 2012; accepted December 3, 2012; published OnlineFirst January 4, 2013.

References

1. Finnish Cancer Registry. Cancer statistics; 2012. [updated 2012 Nov 13]. Available from: www.cancerregistry.fi.
2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78–85.
3. Baker SG, Lichtenstein P, Kaprio J, Holm N. Genetic susceptibility to prostate, breast, and colorectal cancer among Nordic twins. Biometrics 2005;61:55–63.
4. Varghese JS, Easton DF. Genome-wide association studies in common cancers–what have we learnt? Curr Opin Genet Dev 2010;20: 201–9.
5. Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. Hum Mol Genet 2011;20:3867–75.
6. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, et al. Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. Nat Genet 2002;30:181–4.
7. Kote-Jarai Z, Olama AA, Giles GG, Severi G, Schleutker J, Weischer M, et al. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nat Genet 2011;43:785–91.
8. Gillanders EM, Xu J, Chang BL, Lange EM, Wiklund F, Bailey-Wilson JE, et al. Combined genome-wide scan for prostate cancer susceptibility genes. J Natl Cancer Inst 2004;96:1240–7.
9. Xu J, Dimitrov L, Chang BL, Adams TS, Turner AR, Meyers DA, et al. A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. Am J Hum Genet 2005;77:219–29.
10. Cropp CD, Simpson CL, Wahlfors T, Ha N, George A, Jones MS, et al. Genome-wide linkage scan for prostate cancer susceptibility in Finland: evidence for a novel locus on 2q37.3 and confirmation of signal on 17q21-q22. Int J Cancer 2011;129:2400–7.
11. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, et al. Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med 2012;366:141–9.
12. Krumlauf R. Hox genes in vertebrate development. Cell 1994;78: 191–201.
13. Huang L, Pu Y, Hepps D, Danielpour D, Prins GS. Posterior Hox gene expression and differential androgen regulation in the developing and adult rat prostate lobes. Endocrinology 2007;148:1235–45.
14. Kim YR, Oh KJ, Park RY, Xuan NT, Kang TW, Kwon DD, et al. HOXB13 promotes androgen independent growth of LNCaP prostate cancer cells by the activation of E2F signaling. Mol Cancer 2010;9:124.
15. Norris JD, Chang CY, Wittmann BM, Kunder RS, Cui H, Fan D, et al. The homeodomain protein HOXB13 regulates the cellular response to androgens. Mol Cell 2009;36:405–16.
16. Ghoshal K, Motiwala T, Claus R, Yan P, Kutay H, Datta J, et al. HOXB13, a target of DNMT3B, is methylated at an upstream CpG island, and functions as a tumor suppressor in primary colorectal tumors. PLoS ONE 2010;5:e10338.
17. Jerevall PL, Brommesson S, Strand C, Gruvberger-Saal S, Malmstrom P, Nordenskjold B, et al. Exploring the two-gene ratio in breast cancer—independent roles for HOXB13 and IL17BR in prediction of clinical outcome. Breast Cancer Res Treat 2008;107:225–34.
18. Jerevall PL, Jansson A, Fornander T, Skoog L, Nordenskjold B, Stal O. Predictive relevance of HOXB13 protein expression for tamoxifen benefit in breast cancer. Breast Cancer Res 2010;12:R53.
19. Xu J, Lange EM, Lu L, Zheng SL, Wang Z, Thibodeau SN, et al. HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). Hum Genet 2013;132:5–14.
20. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med 2009;360:1320–8.
21. Carter BS, Beaty TH, Steinberg GD, Childs B, Walsh PC. Mendelian inheritance of familial prostate cancer. Proc Natl Acad Sci U S A 1992;89:3367–71.
22. Schleutker J, Matikainen M, Smith J, Koivisto P, Baffoe-Bonnie A, Kainu T, et al. A genetic epidemiological study of hereditary prostate cancer (HPC) in Finland: frequent HPCX linkage in families with late-onset disease. Clin Cancer Res 2000;6:4810–5.
23. Kuusisto KM, Bebel A, Vihinen M, Schleutker J, Sallinen SL. Screening for BRCA1, BRCA2, CHEK2, PALB2, BRIP1, RAD50, and CDH1 mutations in high-risk Finnish BRCA1/2-founder mutation-negative breast and/or ovarian cancer individuals. Breast Cancer Res 2011;13: R20.
24. Syrjakoski K, Vahteristo P, Eerola H, Tamminen A, Kivinummi K, Sarantaus L, et al. Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients. J Natl Cancer Inst 2000;92:1529–31.
25. Eerola H, Blomqvist C, Pukkala E, Pyrhonen S, Nevanlinna H. Familial breast cancer in southern Finland: how prevalent are breast cancer families and can we trust the family history reported by patients? Eur J Cancer 2000;36:1143–8.
26. Fagerholm R, Hofstetter B, Tommiska J, Aaltonen K, Vrtel R, Syrjakoski K, et al. NAD(P)H:Quinone oxidoreductase 1 NQO1*2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer. Nat Genet 2008;40:844–53.
27. Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomaki P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 1998;338:1481–7.
28. Salovaara R, Loukola A, Kristo P, Kaariainen H, Ahtola H, Eskelinen M, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. J Clin Oncol 2000;18: 2193–200.
29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
30. Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 2012;33:1166–74.
31. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 2009;9:51.
32. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59: 467–75.

Published OnlineFirst January 4, 2013; DOI: 10.1158/1055-9965.EPI-12-1000-T

33. Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics 2008;9(Suppl 2):S6.

34. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006;62:1125–32.

35. Huang LT, Gromiha MM, Ho SY. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 2007;23:1292–3.

36. Peltonen L, Jalanko A, Varilo T. Molecular genetics of the Finnish disease heritage. Hum Mol Genet 1999;8:1913–23.

37. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorius H, Bedoya G, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 2006;38:556–60.

38. Sarantaus L, Huusko P, Eerola H, Launonen V, Vehmanen P, Rapakko K, et al. Multiple founder effects and geographical clustering of BRCA1 and BRCA2 families in Finland. Eur J Hum Genet 2000;8:757–63.

39. Lynch HT, Boland CR, Gong G, Shaw TG, Lynch PM, Fodde R, et al. Phenotypic and genotypic heterogeneity in the Lynch syndrome: diagnostic, surveillance and management implications. Eur J Hum Genet 2006;14:390–402.

40. Kestila M, Ikonen E, Lehesjoki AE. [Finnish disease heritage]. Duodecim 2010;126:2311–20.

41. Seppala EH, Ikonen T, Mononen N, Autio V, Rokman A, Matikainen MP, et al. CHEK2 variants associate with hereditary prostate cancer. Br J Cancer 2003;89:1966–70.

42. Cybulski C, Wokolorczyk D, Huzarski T, Byrski T, Gronwald J, Gorski B, et al. A large germline deletion in the Chek2 kinase gene is associated with an increased risk of prostate cancer. J Med Genet 2006;43:863–6.

43. Tischkowitz MD, Yilmaz A, Chen LQ, Karyadi DM, Novak D, Kirchhoff T, et al. Identification and characterization of novel SNPs in CHEK2 in Ashkenazi Jewish men with prostate cancer. Cancer Lett 2008;270: 173–80.

44. Gronwald J, Cybulski C, Piesiak W, Suchy J, Huzarski T, Byrski T, et al. Cancer risks in first-degree relatives of CHEK2 mutation carriers: effects of mutation type and cancer site in proband. Br J Cancer 2009;100:1508–12.

45. CHEK2 Breast Cancer Case–Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet 2004;74:1175–82.

46. Iniesta MD, Gorin MA, Chien LC, Thomas SM, Milliron KJ, Douglas JA, et al. Absence of CHEK2*1100delC mutation in families with hereditary breast cancer in North America. Cancer Genet Cytogenet 2010;202: 136–40.

47. Armenian HK, Lilienfeld AM, Diamond EL, Bross ID. Relation between benign prostatic hyperplasia and cancer of the prostate. A prospective and retrospective study. Lancet 1974;2:115–7.

48. Orsted DD, Bojesen SE, Nielsen SF, Nordestgaard BG. Association of clinical benign prostate hyperplasia with prostate cancer incidence and mortality revisited: a nationwide cohort study of 3,009,258 men. Eur Urol 2011;60:691–8.

49. Berry SJ, Coffey DS, Walsh PC, Ewing LL. The development of human benign prostatic hyperplasia with age. J Urol 1984;132:474–9.

IJC
International Journal of Cancer

Fine-mapping the 2q37 and 17q11.2-q22 loci for novel genes and sequence variants associated with a genetic predisposition to prostate cancer

Virpi H. Laitinen[1], Tommi Rantapero[1], Daniel Fischer[2], Elisa M. Vuorinen[1], Teuvo L.J. Tammela[3], PRACTICAL Consortium, Tiina Wahlfors[1] and Johanna Schleutker[1,4]

[1] BioMediTech, University of Tampere and Fimlab Laboratories, FI-33520, Tampere, Finland
[2] School of Health Sciences, University of Tampere, FI-33014 Tampere, Finland
[3] Department of Urology, Tampere University Hospital and Medical School, University of Tampere, FI-33520 Tampere, Finland
[4] Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, FI-20014 Turku, Finland

Key words: prostate cancer risk, genetic predisposition, susceptibility loci, 2q37, 17q11.2-q22

Abbreviations: AR: Androgen Receptor; ChIP-seq: Chromatin Immunoprecipitation Combined with Massively Parallel DNA Sequencing; CI: Confidence Interval; DB: Database; DE: Differentially Expressed (gene); eQTL: Expression Quantitative Trait Locus; GWAS: Genome Wide Association Study; HPC: Hereditary Prostate Cancer; HWE: Hardy-Weinberg Equilibrium; Indel: Insertion/Deletion Polymorphism; LD: Linkage Disequilibrium; LincRNA: Large Intergenic Non-Coding RNA; LNCaP: Androgen-Sensitive Human Prostate Adenocarcinoma Cell Line Derived From Lymph Node Metastasis; MAF: Minor Allele Frequency; OR: Odds Ratio; PrCa: Prostate Cancer; PSA: Prostate Specific Antigen; PWM: Position Weight Matrix; RNA-seq: Massively Parallel RNA Sequencing; SNP: Single-Nucleotide Polymorphism; SNV: Single-Nucleotide Variant; TF: Transcription Factor; TSS: Transcription Start Site; UTR: Untranslated Region; VCP: Variant-Calling Pipeline; QC: Quality Control

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

The genotyping of variants with Sequenom was performed by the Technology Centre, Institute of Molecular Medicine (FIMM), University of Helsinki, Finland

The PRACTICAL Consortium: Rosalind Eeles: The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, United Kingdom. Royal Marsden NHS Foundation Trust, Fulham and Sutton, London and Surrey, United Kingdom. Doug Easton: Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, United Kingdom. Kenneth Muir: University of Warwick, Coventry, United Kingdom. Graham Giles: Cancer Epidemiology Centre, The Cancer Council Victoria, 1 Rathdowne street, Carlton Victoria, Australia. Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Victoria, Australia. Fredrik Wiklund and Henrik Grönberg: Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden. Christopher Haiman: Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA. Johanna Schleutker: Department of Medical Biochemistry and Genetics, University of Turku, Turku, Finland. BioMediTech, University of Tampere and FimLab Laboratories, Tampere, Finland. Maren Weischer: Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, Herlev Ringvej 75, DK-2730 Herlev, Denmark. Ruth C. Travis: Nuffield Department of Clinical Medicine, Cancer Epidemiology Unit, University of Oxford, Oxford, United Kingdom. David Neal: Surgical Oncology (Uro-Oncology: S4), University of Cambridge, Box 279, Addenbrooke's Hospital, Hills Road, Cambridge, United Kingdom and Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Cambridge, United Kingdom. Paul Pharoah: Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, United Kingdom. Kay-Tee Khaw: Cambridge Institute of Public Health, University of Cambridge, Forvie Site, Robinson Way, Cambridge CB2 0SR, United Kingdom. Janet L. Stanford: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA. Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA. William J. Blot: International Epidemiology Institute, 1455 Research Blvd., Suite 550, Rockville, MD. Stephen Thibodeau: Mayo Clinic, Rochester, MN. Christiane Maier: Department of Urology, University Hospital Ulm, Germany. Institute of Human Genetics University Hospital Ulm, Germany. Adam S. Kibel: Brigham and Women's Hospital/Dana-Farber Cancer Institute, 45 Francis Street, ASB II-3, Boston, MA. Washington University, St Louis, MO. Cezary Cybulski: Department of Genetics and Pathology, International Hereditary Cancer Center, Pomeranian Medical University, Szczecin, Poland. Lisa Cannon-Albright: Division of Genetic Epidemiology, Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT. Hermann Brenner: Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg Germany. Jong Park: Division of Cancer Prevention and Control, H. Lee Moffitt Cancer Center, 12902 Magnolia Dr., Tampa, FL. Radka Kaneva: Molecular Medicine Center and Department of Medical Chemistry and Biochemistry, Medical University—Sofia, 2 Zdrave St, 1431 Sofia, Bulgaria. Jyotsna Batra: Australian Prostate Cancer Research Centre-Qld, Institute of Health and Biomedical Innovation and Schools of Life Science and Public Health, Queensland University of Technology, Brisbane, Australia. Manuel R. Teixeira: Department of Genetics, Portuguese Oncology Institute, Porto, Portugal and Biomedical Sciences Institute (ICBAS), Porto University, Porto, Portugal. Zsofia Kote-Jarai: The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, United Kingdom. Ali Amin Al Olama and Sara Benlloch: University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, United Kingdom.

Int. J. Cancer: **136**, 2316–2327 (2015) © 2014 The Authors. Published by Wiley Periodicals, Inc. on behalf of UICC

The 2q37 and 17q12-q22 loci are linked to an increased prostate cancer (PrCa) risk. No candidate gene has been localized at 2q37 and the *HOXB13* variant G84E only partially explains the linkage to 17q21-q22 observed in Finland. We screened these regions by targeted DNA sequencing to search for cancer-associated variants. Altogether, four novel susceptibility alleles were identified. Two *ZNF652* (17q21.3) variants, rs116890317 and rs79670217, increased the risk of both sporadic and hereditary PrCa (rs116890317: OR = 3.3–7.8, *p* = 0.003–3.3 × 10^{-5}; rs79670217: OR = 1.6–1.9, *p* = 0.002–0.009). The *HDAC4* (2q37.2) variant rs73000144 (OR = 14.6, *p* = 0.018) and the *EFCAB13* (17q21.3) variant rs118004742 (OR = 1.8, *p* = 0.048) were overrepresented in patients with familial PrCa. To map the variants within 2q37 and 17q11.2-q22 that may regulate PrCa-associated genes, we combined DNA sequencing results with transcriptome data obtained by RNA sequencing. This expression quantitative trait locus (eQTL) analysis identified 272 single-nucleotide polymorphisms (SNPs) possibly regulating six genes that were differentially expressed between cases and controls. In a modified approach, prefiltered PrCa-associated SNPs were exploited and interestingly, a novel eQTL targeting *ZNF652* was identified. The novel variants identified in this study could be utilized for PrCa risk assessment, and they further validate the suggested role of *ZNF652* as a PrCa candidate gene. The regulatory regions discovered by eQTL mapping increase our understanding of the relationship between regulation of gene expression and susceptibility to PrCa and provide a valuable starting point for future functional research.

What's new?
Prostate cancer runs in families, but its heritability isn't completely explained by the genetic variants identified to date. In this paper, the authors delve deeper into two loci that have been linked to prostate cancer. Sequencing data revealed four new alleles within these loci that correlate with increased prostate cancer risk. The authors then used the eQTL mapping technique to identify six genes which may be regulated by variants within these two loci, genes which had not previously been associated with prostate cancer.

A large proportion of familial prostate cancer (PrCa) cases can be explained by genetic risk factors.[1] Despite extensive research, the identification of these factors has proven challenging. In Finland, mutations in hereditary prostate cancer (HPC) risk genes are relatively rare, with the exception of the *HOXB13* G84E mutation,[2] which is present in 8.4% of familial PrCa cases and has been significantly associated with an increased PrCa risk in unselected cases.[3]

The involvement of chromosomal regions 2q37 and 17q12-q22 with PrCa has been previously reported in numerous linkage[4–6] and genome-wide association studies (GWASs).[7,8] Cropp *et al.*[9] performed a genome-wide linkage scan of 69 Finnish high-risk HPC families and in the dominant model, the loci on 2q37.3 and 17q21-q22 exhibited the strongest linkage signals. No known PrCa candidate gene resides on 2q37.3, and as demonstrated in our earlier study, the *HOXB13* G84E mutation only partially explains the observed linkage to 17q21-q22.[3]

Here, we performed targeted resequencing that covered the linkage peaks on 2q37 and 17q11.2-q22. The sequence data were filtered to identify the variants within genes predicted to be involved in PrCa predisposition. These variants were validated in Finnish HPC families and in unselected PrCa patients by Sequenom genotyping, and several novel variants were discovered that were significantly associated with PrCa. To study the impact of single-nucleotide polymorphisms (SNPs) on the regulation of gene expression within the two linked regions, we performed transcriptome sequencing followed by expression quantitative trait loci (eQTL) mapping. eQTLs are known to modify the penetrance of rare

Grant sponsor: Academy of Finland; **Grant number:** 251074; **Grant sponsor:** The Finnish Cancer Organisations, the Sigrid Juselius Foundation and the Competitive State Research Financing of the Expert Responsibility Area of Tampere University Hospital; **Grant number:** X51003; **Grant sponsor:** European Commission's Seventh Framework Programme (The PRACTICAL consortium); **Grant number:** HEALTH-F2–2009-223175; **Grant sponsor:** Cancer Research UK; **Grant numbers:** C5047/A7357, C1287/A10118, C5047/A3354, C5047/A10692 and C16913/A6135; **Grant sponsor:** The National Institutes of Health (Cancer Post-Cancer GWAS initiative); **Grant number:** 1 U19 CA 148537-01
DOI: 10.1002/ijc.29276
History: Received 17 Apr 2014; Accepted 1 Oct 2014; Online 21 Oct 2014

Correspondence to: Johanna Schleutker, Medical Biochemistry and Genetics, Institute of Biomedicine, Kiinamyllynkatu 10, University of Turku, FI-20014 Turku, Finland. Tel.: +358-2-3337453, Fax: +358-2-2301280, E-mail: Johanna.Schleutker@utu.fi

deleterious variants and therefore likely contribute to genetic predisposition to complex diseases. New information was obtained on several genes as well as their regulatory elements that generated fresh insights into PrCa susceptibility, especially in HPC.

Material and Methods

All of the subjects were of Finnish origin. The samples were collected with written and signed informed consent. The cancer diagnoses were confirmed using medical records and the annual update from the Finnish Cancer Registry. The project was approved by the local research ethics committee at Pirkanmaa Hospital District and by the National Supervisory Authority for Welfare and Health.

Targeted resequencing of 2q37 and 17q11.2-q22

Based on the linkage analysis results from Cropp *et al.*,[9] 63 PrCa patients and five unaffected individuals belonging to 21 Finnish high-risk HPC families[10] were selected for targeted resequencing of the 2q37 and 17q11.2-q22 regions (Supporting Information Table S1). Each family had at least three first- or second-degree relatives diagnosed with PrCa. Paired-end next generation sequencing was performed at the Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki. The sequenced fragments spanned approximately 6.8 Mb for chromosome 2q and 21.6 Mb for 17q. The target regions were captured using SeqCap EZ Choice array probes (Roche NimbleGen, Madison, WI) and were sequenced on a Genome Analyzer IIx (Illumina, San Diego, CA) following the manufacturer's protocol. The read alignment and variant calling were performed according to FIMM's Variant-Calling Pipeline (VCP).[11]

Bioinformatics workflow for variant characterization

A schematic overview of our bioinformatics workflow is shown in Figure 1. Only those variants that were present in all the affected family members were selected for subsequent analysis. The variants were annotated using Ensembl V65 gene set retrieved from the UCSC Genome Browser.[12] The phenotypic effects of the variants were studied with three *in silico* pathogenicity prediction programs. MutationTaster[13] classifies single-nucleotide variants (SNVs) and small insertion/deletion polymorphisms (indels) as polymorphic or pathogenic. PolyPhen-2[14] and PON-P[15] only predict the effects of nonsynonymous SNVs that result in amino acid replacement. PolyPhen-2 classifies the variants as benign, possibly pathogenic or probably pathogenic, whereas PON-P defines them as neutral, unclassified or pathogenic. Variants categorized as pathogenic by at least one tolerance predictor were defined as pathogenic. In addition, minor allele frequencies (MAFs) were obtained from the dbSNP database and information on known PrCa-associated genes was retrieved from the COSMIC[16] and DDPC[17] databases. Pathway data were gathered from Pathway Commons,[18] KEGG[19] and WikiPathways[20] and Gene Ontology data were retrieved from Ensembl BioMart v.65.[21] Higher priority was assigned to rare variants (MAF <0.05), variants located in genes previously linked to PrCa, and variants located in genes functionally similar to PrCa-associated genes.

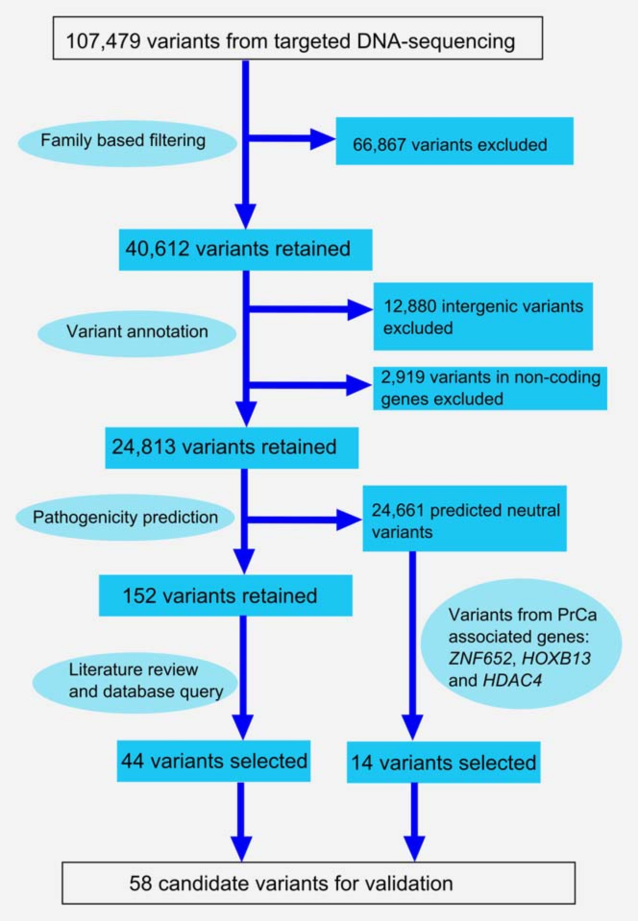

107,479 variants from targeted DNA-sequencing
Family based filtering
66,867 variants excluded
40,612 variants retained
Variant annotation
12,880 intergenic variants excluded
2,919 variants in non-coding genes excluded
24,813 variants retained
Pathogenicity prediction
24,661 predicted neutral variants
152 variants retained
Variants from PrCa associated genes: *ZNF652*, *HOXB13* and *HDAC4*
Literature review and database query
44 variants selected
14 variants selected
58 candidate variants for validation

Figure 1. A flowchart describing the variant characterization pipeline. The targeted resequencing of 2q37 and 17q11.2-q22 from 68 Finnish HPC family members produced a total of 107,479 unique sequence variants. Family-based filtering excluded 66,867 variants that did not cosegregate with affection status. Annotation enabled the selection of 24,813 variants that were located within protein-coding genes. Pathogenicity predictions were performed *in silico* using MutationTaster, PolyPhen-2 and PON-P. As a result, the number of candidate variants was reduced to 152. The final filtering step exploited diverse information on genes and variants as well as gene ontology and pathway data stored in several public databases. In addition, select *HDAC4, ZNF652* and *HOXB13* variants, which were predicted to be nonpathogenic, were included in the validation because these genes have been associated with PrCa in previous studies.

Validation of predicted PrCa-associated variants with Sequenom

After filtering, 58 variants in 35 target genes (listed in Supporting Information Tables S2–S4) were selected for validation which was performed on germline DNA from 2,216 subjects, including 1,293 cases and 923 population controls. The majority of the cases (1,105 individuals) represented unselected PrCa patients from the Pirkanmaa Hospital District, Tampere, Finland. In addition, 188 index cases from Finnish HPC families[10] were included in the study. The control DNA samples from anonymous male blood donors were provided by the Finnish Red Cross Blood Transfusion Service. Genotyping was performed at the Technology Centre, FIMM using the Sequenom MassARRAY system and iPLEX Gold assays (Sequenom, San Diego, CA). Genotyping reactions were performed with 20 ng of dried genomic DNA according to manufacturer's recommendations and with their reagents. The genotypes were called using TyperAnalyzer software (Sequenom). For quality control (QC) reasons, the genotype calls were also checked manually. Genotyping quality was examined using a detailed QC procedure that included success rate checks, duplicated samples and water controls.

Statistical and bioinformatic analyses of the validated variants

Association and Hardy-Weinberg equilibrium (HWE) tests were performed using PLINK.[22] The p value threshold for the HWE test was set to 0.05. Samples with low genotyping frequencies (<0.80) were excluded from the association analysis. The statistical significance of the association was evaluated using a two-sided Fisher's exact test. Odds ratios (OR) were calculated using PLINK with option — fisher. No further model adjustments for confounding factors were made. ENCODE information[23] for noncoding variants was retrieved from the Regulome database (RegulomeDB).[24] The linkage disequilibrium (LD) analysis of the statistically significant variants is described in Supplementary Methods.

Genotyping of the top four candidate variants in Finnish HPC families

Four variants were chosen for segregation analysis in Finnish HPC families based on a strong association with PrCa, a high OR value and/or predicted pathogenicity. The cosegregation of rs116890317 and rs79670217 in *ZNF652* (RefSeq NM_001145365), rs73000144 in *HDAC4* (RefSeq NM_006037) and rs118004742 in *EFCAB13* (RefSeq NM_152347) with affection status was determined in 41 families whose index cases were mutation-positive in the Sequenom validation. For these families, DNA samples were available from 243 PrCa cases and 204 healthy family members. The variants were genotyped in two to 17 (median: seven) individuals per family by Sanger sequencing.

RNA extraction and sequencing

Peripheral blood samples collected in PAXgene® Blood RNA Tubes (PreAnalytiX GmbH, Switzerland) were available from 84 PrCa patients and 15 healthy male relatives belonging to 31 Finnish HPC families. These included 11 families from the targeted resequencing step (Supporting Information Table S1) and additional 20 high-risk families.[10] Total RNA was purified with MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit (Ambion®/Life Technologies, Carlsbad, CA) and with a PAXgene Blood miRNA Kit (PreAnalytiX GmbH). RNA integrity and quality were analyzed using the Agilent 2100 Bioanalyzer and the Agilent RNA 6000 Nano Kit (Agilent Technologies, Santa Clara, CA). The massively parallel paired-end RNA sequencing was performed at Beijing Genomics Institute (BGI Hong Kong Co., Tai Po, Hong Kong) using an Illumina HiSeq2000 sequencing platform (Illumina).

RNA sequencing data analysis

On average, RNA sequencing produced 45 million reads per sample. The QC check was performed using fastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). The reads were aligned with Tophat2[25] using GRCh37/hg19 as the reference genome. The read counts for the genes were determined using HTSeq (http://www-huber.embl.de/users/anders/HTSeq/). The raw read counts were transformed into comparable expression values via normalization using the DESeq package for R[26] and the genes with very low or no expression (normalized read counts of <20) were removed. A differential gene expression analysis was then performed using a two-sided Mann–Whitney test with a p value cutoff of 0.05.

eQTL mapping and data analysis

The eQTL analysis was based on the RNA-seq data and on the SNP genotypes obtained from targeted DNA sequencing. This data existed for 19 samples at 2q37 and for 17 samples at 17q11.2-q22. In total, 54,919 SNPs (average 6,865 per gene, see Supporting Information Table S5 for details) were tested for association with their candidate target genes. Only genes with differential expression (DE) patterns between health status groups were included in the eQTL analysis, to increase the probability that found SNP-gene associations also link PrCa with a certain SNP genotype. The eQTL mapping was applied on 2q37 and 17q11.2-q22 to identify *cis*-regulated genes. SNPs associated in *cis* were defined as variants located within 1 Mb up- or downstream of the gene under study. The significance level for SNP-gene associations was set to $p \leq 0.005$. A multiple testing adjustment was omitted because of the large number of tested SNPs and the nature of the permutation type tests, acknowledging that this resulted in compromised resolution.

A modified *cis*-eQTL approach was also utilized, wherein a large genotype dataset from the iCOGS study[27] was used to preidentify possible PrCa-associated SNPs for 2,824 unselected Finnish PrCa patients and 2,440 controls. Here, Fisher's exact test with a modest significance level of 0.005 was used to study the association. Significant iCOGS variants that were also observed in the targeted DNA sequencing data were then selected for eQTL analysis, which was restricted to the fine-mapped regions. Additional details for the eQTL analysis are presented in Supplementary Methods.

RegulomeDB was used to annotate and assess the regulatory potential of the detected eQTLs.[24] The ENCODE

Table 1. Variants significantly associated with prostate cancer based on a comparison of familial cases ($n = 186$) and controls ($n = 914$)

SNP Id	Function	Gene	Chr	Min/Maj	F_A/F_U (%)	*p* value	OR (95% CI)	Pathogenicity prediction
rs116890317	Intronic	*ZNF652*	17	A/T	2.96/0.39	**3.3×10^{-5}**	7.8 (3.0–20.3)	Polymorphism/–/–
rs79670217	Intronic	*ZNF652*	17	G/T	6.65/3.56	**0.009**	1.9 (1.2–3.1)	Polymorphism/–/–
rs10554930	Intronic	*HOXB3*	17	−ACA/ACA	27.5/21.3	**0.010**	1.4 (1.1–1.8)	Pathogenic/–/–
rs35384813	5′-UTR	*HOXB3*	17	+T/−	26.7/20.8	**0.013**	1.4 (1.1–1.8)	Pathogenic/–/–
rs73000144	Missense	*HDAC4*	2	T/C	0.80/0.06	**0.018**	14.6 (1.5–140.2)	Polymorphism/benign/neutral
rs13411615[1]	Near gene 5′	*MYEOV2*	2	C/A	52.1/45.6	**0.023**	1.3 (1.0–1.6)	Polymorphism/–/–
rs9899142	Intronic	*HOXB13*	17	T/C	11.2/15.6	**0.031**	0.7 (0.5–1.0)	Polymorphism/–/–
rs118004742	Nonsense	*EFCAB13*	17	G/T	4.79/2.73	**0.048**	1.8 (1.0–3.1)	Pathogenic/–/–
rs142044482	3′-UTR	*ZNF652*	17	+A/−	2.94/1.59	0.087	1.9 (0.9–3.8)	Polymorphism/–/–
rs140611363[1]	Near gene 5′	*ACACA*	17	−A/A	28.8/31.1	0.421	0.9 (0.7–1.1)	Pathogenic/–/–
rs72828246[1]	Near gene 5′	*ACACA*	17	G/A	28.8/30.9	0.459	0.9 (0.7–1.2)	Pathogenic/benign/neutral
rs13406410[1]	Near gene 5′	*MYEOV2*	2	C/T	47.6/46.8	0.817	1.0 (0.8–1.3)	Pathogenic/–/–
rs61752234	Synonymous	*HDAC4*	2	C/T	7.22/6.83	0.823	1.1 (0.7–1.6)	Polymorphism/–/–

Bold signifies $p < 0.05$.

[1] Variants are in linkage disequilibrium.

Abbreviations: Chr: chromosome; Min: minor allele; Maj: major allele; F_A: frequency of the minor allele in cases; F_U: frequency of the minor allele in controls; OR: odds ratio; CI: confidence interval; pathogenicity prediction results from: MutationTaster/PolyPhen-2/Pon-P.

datasets[23] were retrieved from the UCSC Genome Browser website for visualization purposes using the Table Browser tool.[12] As a general indicator of regulatory potential, we used the dataset that contained enriched DNase hypersensitive sites in 125 cell types. To highlight the regulatory potential of eQTLs in PrCa tissue, we used the LNCaP DNase (wgEncodeAwgDnaseUwDukeLncapUniPk) and LNCaP (Andr) DNase (wgEncodeAwgDnaseUwDukeLncapandrogenUniPk) datasets containing DNase hypersensitive sites in LNCaP cells under normal and androgen-induced conditions, respectively. Transcription factor (TF) binding site data were gathered from the Txn Fac ChIP V3 dataset, which contains ChIP-seq experimental data on 91 cell types and 189 TFs.

Results

Targeted DNA sequencing data analysis

The percentage of mapped reads was 95.0 and 95.7% for the samples sequenced for 2q37 and 17q11.2-q22, respectively. The target coverage was 99.8% for 2q37 and 99.5% for 17q11.2-q22. Correspondingly, the percentage of bases having coverage of 20× or more was 79.9 and 63.4%. The total number of unique variants across all samples discovered by the utilized VCP was 107,479 (Fig. 1). Among the 41 predicted pathogenic variants in 2q37, there were 20 missense SNVs, 16 noncoding SNVs and five indels. Of all 111 predicted pathogenic variants in 17q11.2-q22, two variants were nonsense SNVs, 49 were missense SNVs, 36 were noncoding SNVs and 24 were indels.

PrCa-associated variants identified by Sequenom validation

Following prioritization, a total of 58 variants were selected for validation in a larger sample set (Supporting Information Table S2). In the QC analysis, four variants failed the HWE test ($p < 0.05$), and 20 samples were omitted due to low genotyping frequencies (<0.80). In the case-control association analysis, a total of 13 variants in seven different genes were statistically significantly associated with PrCa ($p < 0.05$; Tables 1 and 2 and Supporting Information Tables S3 and S4). Three variants were located in the *ZNF652* gene at 17q21.3, and the *HDAC4* (2q37.2), *HOXB3* (17q21.3), *ACACA* (17q21) and *MYEOV2* (2q37.3) genes harbored two variants each. A single variant was identified in the *HOXB13* and *EFCAB13* genes at 17q21.3. Only three of these 13 PrCa-associated variants were located within exons, whereas the majority, 10 variants, resided in noncoding regions.

Four of the variants with a statistically significant association with PrCa were present in both the familial and the unselected sample sets. These were rs116890317 and rs79670217 in *ZNF652*, rs10554930 in *HOXB3* and rs13411615 in *MYEOV2*. The two *ZNF652* variants had the strongest association with an increased PrCa risk. rs116890317 had the most significant association with the familial cases (OR = 7.8, 95% CI 3.0–20.3, $p = 3.3 \times 10^{-5}$) and the same variant conferred the highest risk of 3.3 (95% CI 1.4–7.5, $p = 0.003$) among the unselected cases. rs79670217 had the most significant association with PrCa in the unselected sample set ($p = 0.002$) and was the second most significant variant in the familial PrCa patients (OR = 1.9, 95% CI 1.2–3.1, $p = 0.009$; Tables 1 and 2).

The highest OR of 14.6 (95% CI 1.5–140.2, $p = 0.018$) was observed for the *HDAC4* variant rs73000144 (*c.958C>T, p.Val320Ile*) among the familial samples (Table 1). Only three familial PrCa patients (1.6%), seven unselected patients (0.6%) and one control individual (0.1%) carried the minor allele in a heterozygous state, and none of the genotyped individuals were homozygous. rs73000144 was predicted to

Table 2. Variants significantly associated with prostate cancer based on a comparison of unselected cases ($n = 1096$) and controls ($n = 914$)

SNP Id	Function	Gene	Chr	Min/Maj	F_A/F_U (%)	*p* value	OR (95% CI)	Pathogenicity prediction
rs79670217	Intronic	*ZNF652*	17	G/T	5.66/3.56	**0.002**	1.6 (1.2–2.2)	Polymorphism/–/–
rs116890317	Intronic	*ZNF652*	17	A/T	1.27/0.39	**0.003**	3.3 (1.4–7.5)	Polymorphism/–/–
rs13406410[1]	Near gene 5′	*MYEOV2*	2	C/T	51.5/46.8	**0.006**	1.2 (1.1–1.4)	Pathogenic/–/–
rs61752234	Synonymous	*HDAC4*	2	C/T	4.85/6.83	**0.008**	0.7 (0.5–0.9)	Polymorphism/–/–
rs142044482	3′-UTR	*ZNF652*	17	+A/-	0.68/1.59	**0.009**	0.4 (0.2–0.8)	Polymorphism/–/–
rs140611363[1]	Near gene 5′	*ACACA*	17	−A/A	27.9/31.1	**0.032**	0.9 (0.7–1.0)	Pathogenic/–/–
rs10554930	Intronic	*HOXB3*	17	−ACA/ACA	24.1/21.3	**0.034**	1.2 (1.0–1.4)	Pathogenic/–/–
rs13411615[1]	Near gene 5′	*MYEOV2*	2	C/A	49.0/45.6	**0.037**	1.1 (1.0–1.3)	Polymorphism/–/–
rs72828246[1]	Near gene 5′	*ACACA*	17	G/A	28.0/30.9	**0.044**	0.9 (0.8–1.0)	Pathogenic/benign/neutral
rs35384813	5′-UTR	*HOXB3*	17	+T/–	23.2/20.8	0.073	1.1 (1.0–1.3)	Pathogenic/–/–
rs73000144	Missense	*HDAC4*	2	T/C	0.33/0.06	0.078	5.9 (0.7–47.9)	Polymorphism/benign/neutral
rs118004742	Nonsense	*EFCAB13*	17	G/T	3.0/2.7	0.637	1.1 (0.8–1.6)	Pathogenic/–/–
rs9899142	Intronic	*HOXB13*	17	T/C	16.1/15.6	0.665	1.0 (0.9–1.2)	Polymorphism/–/–

Bold signifies $p < 0.05$.
[1] Variants are in linkage disequilibrium.
Abbreviations: Chr: chromosome; Min: minor allele; Maj: major allele; F_A: frequency of the minor allele in cases; F_U: frequency of the minor allele in controls; OR: odds ratio; CI: confidence interval; pathogenicity prediction results from: MutationTaster/PolyPhen-2/Pon-P.

be benign or neutral by all three *in silico* pathogenicity prediction algorithms (Supporting Information Table S2).

The rs118004742 nonsense mutation (*c.1638T>G, p.Tyr546Ter*) in the *EFCAB13* gene was predicted to be pathogenic by MutationTaster (Supporting Information Table S2). Three familial cases (1.6%) were homozygous for the minor allele. There were 12 heterozygotes among the familial index cases (6.5%) and 66 among the unselected cases (6.0%). A statistically significant association between rs118004742 and PrCa was only observed for the familial patients (Table 1). The OR of 1.8 (95% CI 1.0–3.1) suggested an increased risk of HPC. rs118004742 carriers in the unselected sample set did not have an increased cancer risk (OR = 1.1, 95% CI 0.8–1.6, $p = 0.637$; Supporting Information Table S4).

Two common noncoding variants in the *HOXB3* gene, rs10554930 and rs35384813, had a moderate effect on PrCa risk, with ORs ranging from 1.2 to 1.4 (Tables 1 and 2). MutationTaster predicted both of these variants to be pathogenic (Supporting Information Table S2). For five variants, the ORs were < 1.0, indicating a modulatory role in PrCa predisposition. These variants were located near or within the *ZNF652, HDAC4, HOXB13* and *ACACA* genes (Tables 1 and 2). According to the RegulomeDB, three of the 13 statistically significant variants were likely to affect protein binding: rs9899142 in *HOXB13* (Regulome score of 1f), rs13406410 in *MYEOV2* and rs72828246 in *ACACA* (both having Regulome score of 2b).

In case-case comparisons, none of the identified variants were significantly associated with Gleason score, average age or the serum prostate specific antigen (PSA) level at diagnosis (data not shown). The LD analysis (Supporting Information Fig. S1) revealed that none of our 13 statistically significant variants (Tables 1 and 2) were in linkage disequilibrium with previously reported PrCa-associated variants[27] (see Supplementary Results for details).

Segregation analysis of the top four candidate variants

Altogether, 41 familial index cases out of 188 genotyped by Sequenom carried at least one of the top four candidate variants. Segregation analysis was performed for these 41 HPC families. rs116890317, rs79670217 and rs118004742 were more common among PrCa patients than healthy family members and provided evidence for cosegregation with affection status in 20 families (Supporting Information Tables S6–S8). However, in 15 of these families, unaffected male mutation carriers were also observed. In seven families, all of the unaffected male carriers were young enough (<55 years) to develop PrCa later in life. rs116890317 segregated completely with affection status in one family (Supporting Information Fig. S2*a*), as did rs79670217 (Supporting Information Fig. S2*b*). Complete segregation of rs118004742 was observed in three families (Supporting Information Table S8). The *HDAC4* variant rs73000144 was detected in three families, and approximately one-third of the family members were identified as carriers, irrespective of their health status (Supporting Information Table S9).

Multiple variants were observed in 16 individuals from 14 families. Two families harbored rs116890317, rs79670217 and rs118004742, whereas one family was positive for rs79670217, rs73000144 and rs118004742. In the remaining families, the most common combination detected was rs79670217 together with rs118004742 (six families). Evidence for segregation with affection status was obtained for a maximum of one variant per family.

Cancer Genetics

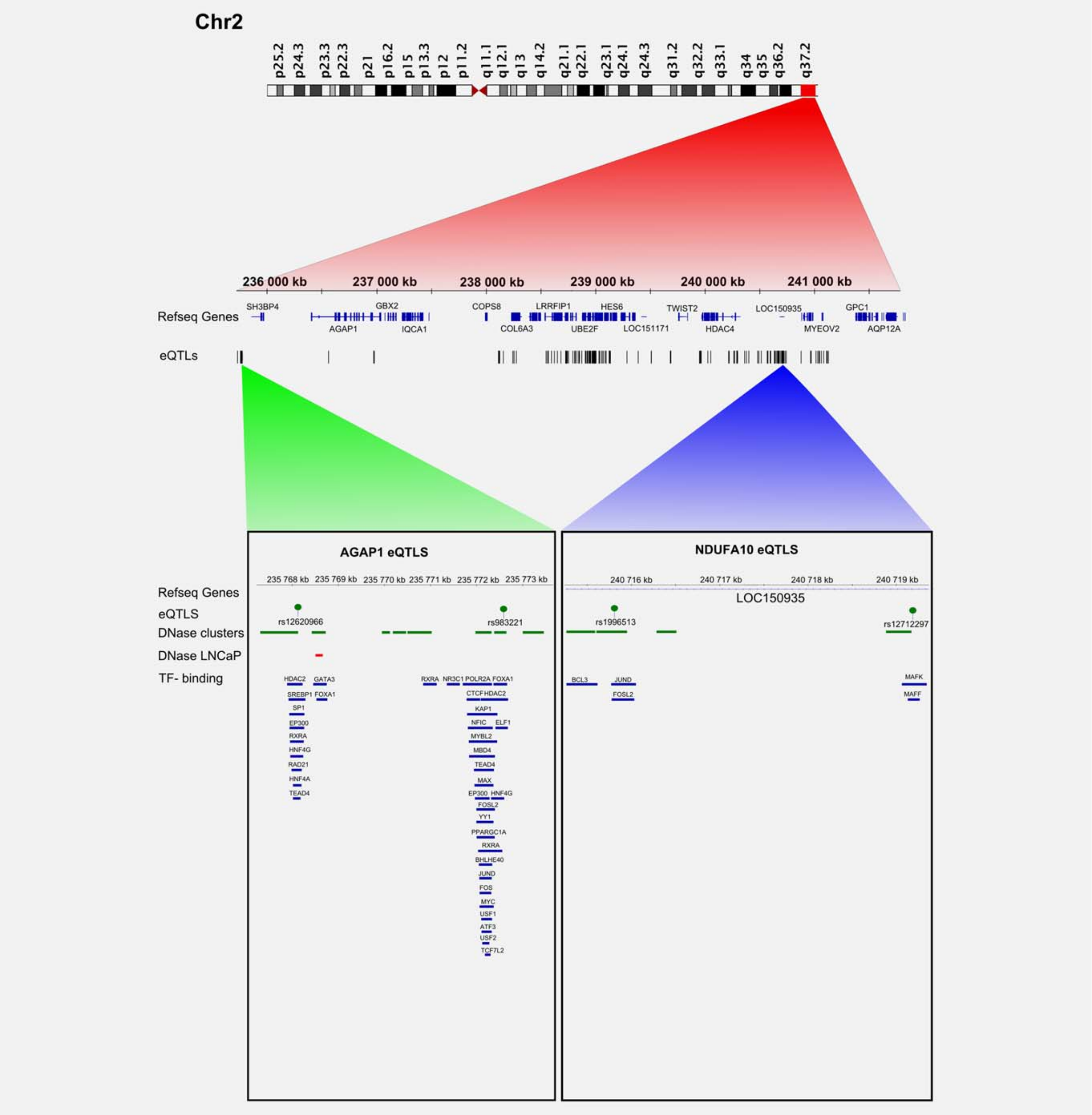


Figure 2. *Cis*-eQTLs targeting differentially expressed genes on chromosome 2. All statistically significant eQTLs are indicated with a track of black bars. Selected eQTLs, rs12620966 and rs983221 (targeting *AGAP1*) and rs1996513 and rs12712297 (targeting *NDUFA10*) are illustrated in more detail. DNaseI hypersensitive sites from the DNase cluster and LNCaP datasets are indicated with green and red rectangles, respectively. Blue rectangles denote TF binding sites.

eQTL mapping results

Differential gene expression analysis revealed three genes (of 173 tested) located at 2q37 and five genes (of 761 tested) at 17q11.2-q22 whose expression levels differed significantly between cases and controls ($p < 0.05$). In the targeted *cis*-eQTL analysis, SNPs within 2 Mb windows were tested for association with each of these eight DE genes (Supporting Information Table S5). Altogether, 272 candidate regulatory SNPs were identified for six DE genes only (Supporting Information Table S10). A vast majority, 237 candidate SNPs potentially regulate the expression of *AGAP1, SCLY* and *NDUFA10* at 2q37 (Fig. 2). The remaining 35 candidate SNPs possibly regulate *TBKBP1, PNPO* and *NAGS* at 17q11.2-q22 (Fig. 3). Based on the ENCODE data, the strongest evidence for regulatory potential was found for rs11650354 on chromosome 17, which targets the *TBKBP1*

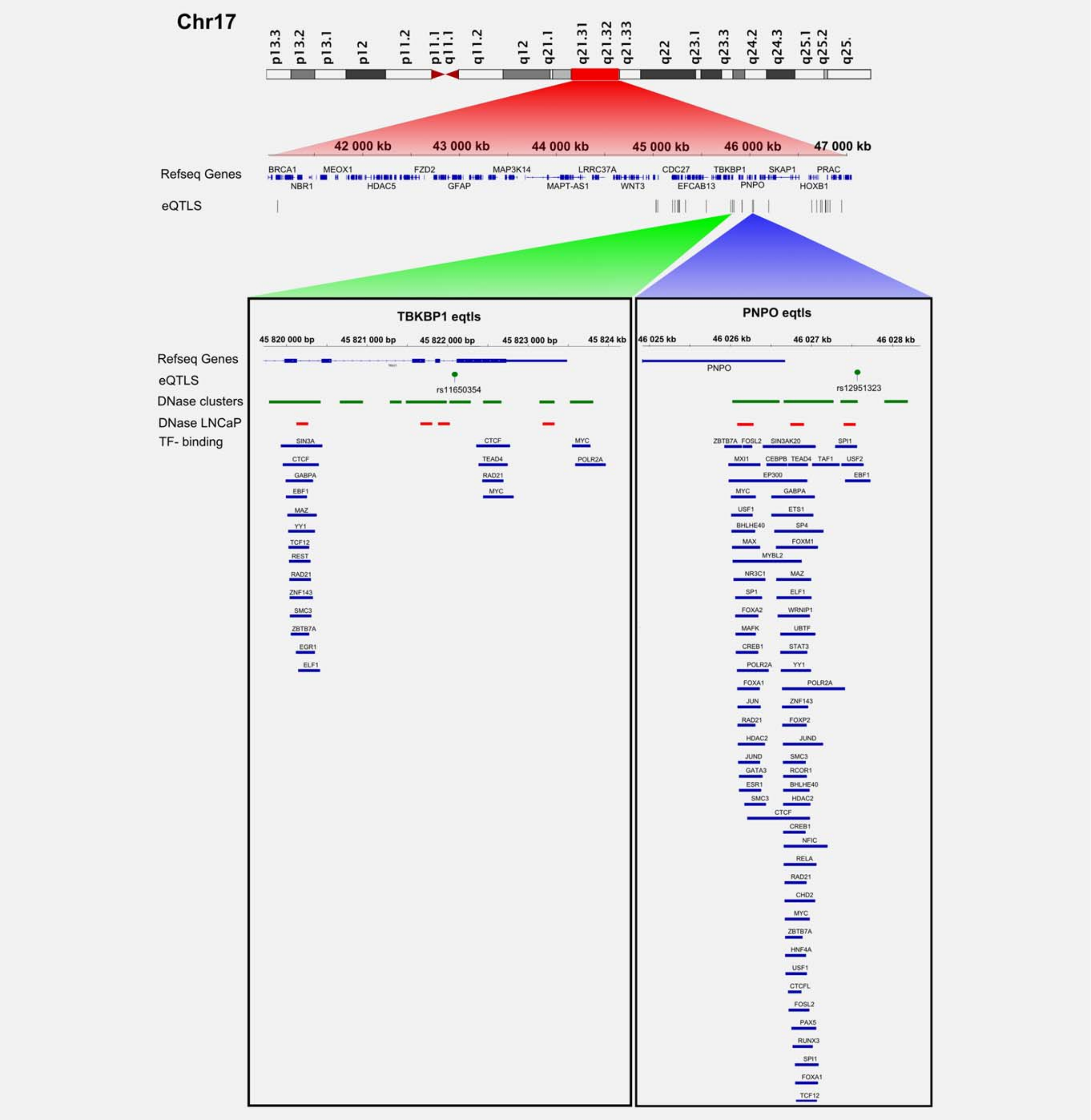


Figure 3. *Cis*-eQTLs targeting differentially expressed genes on chromosome 17. All statistically significant eQTLs are indicated with a track of black bars. Selected eQTLs, rs11650354 (targeting *TBKBP1*) and rs12951323 (targeting *PNPO*) are illustrated in more detail. DNaseI hypersensitive sites from the DNase cluster and LNCaP datasets are indicated with green and red rectangles, respectively. Blue rectangles denote TF binding sites.

gene. This known eQTL overlaps with an open chromatin region (Mcf7 and Gm12892 cell lines) and its role in the regulation of *TBKBP1* expression has been confirmed in a previous study.[28] rs12620966 targeting *AGAP1* on chromosome 2 overlaps with several TF binding sites discovered by ChIP-seq (HepG2 cell line), position weight matrix (PWM) matching and digital DNaseI footprinting studies (Supporting Information Table S10). None of the coding variants that were identified by targeted DNA sequencing and validated by Sequenom were statistically significant eQTLs (data not shown).

The modified *cis*-eQTL analysis was based on 12 SNPs at 2q37 and 22 SNPs at 17q11.2-q22 that were shared between the iCOGS dataset and our set of variants obtained by targeted resequencing. The regulatory potential of these 34 SNPs was evaluated for 144 genes at 2q37 and for 160 genes at 17q11.2-q22. The modified eQTL approach identified only one PrCa-associated candidate eQTL on chromosome 2 and 36 candidate eQTLs on chromosome 17. Selected examples of these eQTLs and their target genes are shown in Supporting Information Table S11. The ENCODE data from RegulomeDB indicated the strongest evidence of regulatory potential for two variants on chromosome 17, rs4796751 and rs4796616, which target the *DHX58, MLX* and *JUP* genes. Both variants have previously been reported as eQTLs targeting *MGC20781* and *NT5C3L*[29] and they overlap with open chromatin regions (in 16 and 17 cell lines, respectively). rs4796616 is also located within a TF binding site (U2OS cell line). Two additional chromosome 17 variants, rs4793943 and rs16941107 were defined as likely to affect gene expression. These variants target the *ZNF652* and *ARL17B* genes, respectively, and overlap with open chromatin regions (in 6 and 42 cell lines, respectively) as well as several TF binding sites (Supporting Information Table S11). Of particular interest was the chromosome 17 variant rs4793976 targeting the *SPOP* gene. Although no data for this eQTL was available in the RegulomeDB, the importance of *SPOP* in PrCa predisposition has been recognized.[30]

Discussion

Prior studies have identified a strong relationship between PrCa and linkage to chromosomal regions 2q37 and 17q11.2-q22. Inspired by the lack of candidate genes and mutations, we resequenced the linkage peaks and confirmed the sequencing results by validating select variants. As the number of variants provided by the VCP was high, their prioritization for validation was critical.

The variants that were statistically significantly associated with PrCa were clustered in two genes on chromosome 2q37, *HDAC4* and *MYEOV2*, and in five genes on chromosome 17q11.2-q22, *ZNF652, HOXB3, HOXB13, EFCAB13* and *ACACA* (Tables 1 and 2). Interestingly, four of these genes, *HDAC4, ZNF652, HOXB3* and *HOXB13* encode TFs. Transcriptional regulation plays an essential role in maintaining normal gene control, and mutations in genes coding for TFs have been identified in PrCa. Examples of commonly occurring alterations include the fusion of *TMPRSS2* with *ERG*, and mutations in genes coding for the forkhead-box family of TFs.[31]

The *ZNF652* gene at 17q21.3 codes for a DNA-binding transcriptional repressor protein with seven zinc finger motifs.[32] Highest expression levels have been detected in normal breast, prostate and pancreas, whereas in primary tumors and cancer cell lines, *ZNF652* expression is generally lower.[32] However, in PrCa, the coexpression of high levels of ZNF652 and the androgen receptor (AR) has been shown to increase the risk of PSA relapse.[33] In addition, the recently characterized ZNF652 DNA binding site was found in the promoters of several genes that are involved in PrCa development and progression.[34] ZNF652 also interacts with CBFA2T3, a putative breast cancer tumor suppressor, which has been shown to enhance the repressor activity of ZNF652.[32]

To date, only a single PrCa-associated risk variant has been identified in the *ZNF652* gene. rs7210100 has been reported to predispose men of African descent to PrCa. The risk allele is present at a low frequency (<1%) in non-African populations.[35] A possible European-specific risk variant, rs11650494, is located in a lincRNA just downstream of the *ZNF652* gene and was recently described by the PRACTICAL Consortium.[27] The present study identified two novel *ZNF652* gene variants, rs116890317 and rs79670217, which were significantly associated with PrCa in both familial and unselected cases. The risk association was particularly apparent in patients with a positive family history of the disease. Correspondingly, both variants showed evidence for at least partial cosegregation with affection status in a substantial portion of Finnish HPC families. Like rs7210100, these two novel variants are located in the first intron of the gene, suggesting that they may play a role in regulating *ZNF652* by affecting splicing events and/or tissue-specific expression.

The *HDAC4* gene at 2q37.2 encodes a well-characterized transcriptional repressor. HDAC4 has been reported to accumulate in the nucleus in hormone-refractory PrCa[36] and to bind to and inhibit the activity of AR by SUMOylation.[37] Here, we determined that the exonic *HDAC4* variant rs73000144 (*c.958C>T*) was significantly associated with familial PrCa (OR = 14.6, 95% CI 1.5–140.2, $p = 0.018$). The variant also had a high OR (=5.8, 95% CI 0.7–47.9) among the unselected cases (Supporting Information Table S4), suggesting an increased cancer risk, but this result was not statistically significant ($p = 0.078$). The pathogenicity of rs73000144 is uncertain. The resulting amino acid change, a substitution of isoleucine for valine (*p.Val320Ile*) is conservative and was not considered pathogenic by any of the *in silico* predictors used (Supporting Information Table S2). The strikingly high OR for the familial sample set, together with the observation that this variant was detected in only three out of 186 index cases from the Finnish HPC families, suggested that rs73000144 may be a private mutation. The importance of private mutations has been emphasized in many diseases, some of which are associated with specific ethnic groups.

The protein encoded by the *EFCAB13* (*EF-hand calcium binding domain 13*) gene at 17q21.3 contains a particular helix-loop-helix domain, the EF-hand, which is required for calcium ion binding. EF-hands are often found in calcium sensor and calcium signal modulator proteins. Ca^{2+} binding triggers a conformational change in the EF-hand motif, which leads to the activation or inactivation of target proteins. Currently, there is no evidence linking EFCAB13 with PrCa. The nonsense mutation rs118004742 in the *EFCAB13* gene introduces a premature stop codon, leading to a significant truncation of the nascent protein. Truncating mutations are generally considered deleterious and, as expected, rs118004742 was predicted pathogenic by MutationTaster (Supporting Information Table S2). The variant segregated completely with affection status in three Finnish mutation-positive HPC families and showed evidence for partial cosegregation in four additional families. In these seven families, the variant was observed in all of the patients but in only half of the genotyped unaffected men (Supporting Information Table S8). It is possible that rs118004742 contributes to hereditary, but not sporadic, disease. Once a more detailed characterization of the EFCAB13 protein function is available, it will be possible to assess the indicative role of *EFCAB13* as a PrCa risk gene more accurately.

Considering the importance of the *HOXB13* variant G84E[2] in familial PrCa predisposition, we compared the families that were positive for the top four SNPs with the existing G84E genotyping data.[3] Interestingly, ten of the 11 families that were positive for the *ZNF652* variant rs116890317 also harbored G84E. In these ten families, 12/21 (57%) of PrCa patients carried both the rs116890317 variant and the *HOXB13* variant G84E. Cosegregation of the *ZNF652* variant rs79670217 (Supporting Information Table S7) and G84E was detected in 6/42 (14%) of affected individuals, and among the 31 PrCa patients carrying the *EFCAB13* variant rs118004742 (Supporting Information Table S8), G84E was identified in only 2 (6%) patients. In addition, one of the three PrCa patients carrying the *HDAC4* variant rs73000144 also carried G84E. The co-occurrence of the *ZNF652* variant rs116890317 with the *HOXB13* variant G84E suggests possible interaction between these two genomic regions and is an interesting issue for future research.

The *HOXB3* gene belongs to the same evolutionarily conserved *HOXB* gene family at 17q21-q22 as *HOXB13*. Recently, HOXB3 overexpression was observed in primary PrCa tissues, predicting poor survival.[38] In our study, two possibly pathogenic *HOXB3* variants were associated with a moderately increased PrCa risk, rs10554930 in both datasets and rs35384813 in the familial sample set only (Tables 1 and 2). rs10554930 is intronic, located ~730 bp upstream of the *HOXB3* transcription start site (TSS), whereas rs35384813 is in the 5′-UTR of the gene. Most variants affecting the expression level of a particular gene are located near the TSS of that gene[29] making it possible that these two variants participate in the regulation of *HOXB3* gene expression.

The ENCODE data supported a possible regulatory role for three of the statistically significant noncoding variants validated by Sequenom. The intronic *HOXB13* variant rs9899142 likely affects the binding of ZNF263, a transcriptional repressor that participates in cell structure maintenance and proliferation.[39] This variant is also a known *cis*-eQTL that regulates the expression of the *SKAP1* gene which has been associated with PrCa-specific mortality.[40] The SNPs rs13406410 and rs72828246 are located near the 5′ ends of the *MYEOV2* and *ACACA* genes, respectively. Both of these variants likely affect the binding of E2F1. This TF plays a central role in DNA damage-induced apoptosis and DNA repair.[41] Recently, a strong correlation between E2F1 and increased expression of NuSAP, a protein that binds DNA to the mitotic spindle, was observed in recurrent PrCa.[42] The minor alleles of rs9899142, rs13406410 and rs72828246 had a low OR and were present at a high frequency in both cases and controls. Nevertheless, according to the common disease–common variant hypothesis, it is possible that the major alleles, rather than the minor alleles, explain a proportion of PrCa susceptibility.

The eQTL mapping enabled us to identify genomic regions that were likely to be regulated by variants in the 2q37 and 17q11.2-q22 loci. A drawback of the eQTL analysis was the use of peripheral blood for RNA-sequencing. However, fresh PrCa tissue is rarely available and, due to the multifocal nature of PrCa, the quality of prostate biopsies may be compromised. Postmortem material, on the other hand, represents expression profiles typical for end-stage disease, whereas our aim was to identify inherited mutations predisposing their carriers to PrCa. Therefore, we consider blood to be a valid starting point for expression profiling of the early changes in PrCa. It will be exciting to see whether future studies confirm our results in another, independent sample set, preferably a collection of PrCa tissue samples.

The traditional eQTL analysis identified six DE genes that were putatively regulated by eQTLs in *cis* (Figs. 2 and 3; Supporting Information Table S10). None of these genes has previously been associated with PrCa. The protein encoded by the *AGAP1* gene is involved in membrane trafficking and cytoskeleton dynamics.[43] SCLY and PNPO participate in metabolic processes, SCLY in the decomposition of L-selenocysteine[44] and PNPO in the biosynthesis of vitamin B6. The adaptor protein encoded by *TBKBP1* plays a role in the TNF-alpha/NF-kappa B signal transduction pathway.[45] NDUFA10 and NAGS are mitochondrial enzymes. NDUFA10, a member of the respiratory chain complex I, is responsible for electron transport.[46] NAGS catalyzes the formation of N-acetylglutamate, an activator of urea cycle enzyme CPSI.[47]

In the modified eQTL analysis, several *cis*-acting variants that were associated with altered gene expression were identified (Supporting Information Table S11). The most interesting finding was the association of rs4793943 with *ZNF652* expression. This interaction may alter the TF function of ZNF652, thereby modulating susceptibility to PrCa. Data from RegulomeDB suggest that rs4793943 may have a more generalized role in transcriptional regulation. It is located within the binding site of ZNF263[39] and it overlaps with HOXA9 and HOXB13 binding motifs. Both of these TFs have been connected with PrCa initiation and progression.[2,48] Furthermore, our data provided suggestive evidence that rs4793976 is an eQTL regulating the expression of SPOP (Supporting Information Table S11). *SPOP*, a putative tumor suppressor gene, is frequently mutated in localized and advanced prostate tumors.[30] *SPOP* mutations are regarded as driver lesions in prostate carcinogenesis[31] and the loss of SPOP expression may contribute to PrCa development.[49]

While interpreting the eQTL results, it is important to recall that the significant DE genes and SNP-gene associations could be identified merely by chance. The number of observed significant test results lies in the same magnitude as the number of expected significant test results, if the null hypothesis would hold for all performed tests. However, the risk of an excess of false positive results was accepted in favor of minimizing the risk of obtaining too many false negative results. Although several of the SNP-gene connections detected in this study achieved statistical significance, this does not necessarily indicate biological significance. Neither is the mechanism of interaction between the individual eQTLs and their target genes currently known. Further validation with independent datasets is required to confirm the significance of the SNP-gene associations identified here.

In conclusion, the present study demonstrated that next-generation sequencing is a valid and reliable approach for identifying novel disease-associated variants and mutations, especially those rare enough to escape the resolution of GWAS. In contrast to imputation and related prediction-based methods, next-generation sequencing methods provide true genotype data with a minimal error rate. The integrated analysis of rare and common variants with gene expression data generated unique knowledge of PrCa-associated variants with effects at the transcriptional level. This study provided a broader view of the causative factors in PrCa, implicating that regulatory variants co-operating with coding variants can modulate the inherited risk for the disease. The findings reported here encourage further research to elucidate the regulatory networks that control PrCa initiation and development.

Acknowledgements

The authors wish to thank all the patients and families who participated in this study. The authors also thank Ms. Riitta Vaalavuo and Ms. Riina Kylätie for technical assistance.

References

1. Baker SG, Lichtenstein P, Kaprio J, et al. Genetic susceptibility to prostate, breast, and colorectal cancer among Nordic twins. *Biometrics* 2005;61: 55–63.
2. Ewing CM, Ray AM, Lange EM, et al. Germline mutations in HOXB13 and prostate-cancer risk. *N Engl J Med* 2012;366:141–9.
3. Laitinen VH, Wahlfors T, Saaristo L, et al. HOXB13 G84E mutation in Finland: population-based analysis of prostate, breast, and colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev* 2013;22:452–60.
4. Xu J, Dimitrov L, Chang BL, et al. A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am J Hum Genet* 2005;77:219–29.
5. Lange EM, Robbins CM, Gillanders EM, et al. Fine-mapping the putative chromosome 17q21–22 prostate cancer susceptibility gene to a 10 cM region based on linkage analysis. *Hum Genet* 2007;121:49–55.
6. Pierce BL, Friedrichsen-Karyadi DM, McIntosh L, et al. Genomic scan of 12 hereditary prostate cancer families having an occurrence of pancreas cancer. *Prostate* 2007;67:410–5.
7. Gudmundsson J, Sulem P, Steinthorsdottir V, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* 2007;39: 977–83.
8. Eeles RA, Kote-Jarai Z, Giles GG, et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* 2008;40:316–21.
9. Cropp CD, Simpson CL, Wahlfors T, et al. Genome-wide linkage scan for prostate cancer susceptibility in Finland: evidence for a novel locus on 2q37.3 and confirmation of signal on 17q21-q22. *Int J Cancer* 2011;129:2400–7.
10. Schleutker J, Matikainen M, Smith J, et al. A genetic epidemiological study of hereditary prostate cancer (HPC) in Finland: frequent HPCX linkage in families with late-onset disease. *Clin Cancer Res* 2000;6:4810–5.
11. Sulonen AM, Ellonen P, Almusa H, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 2011;12:R94,2011–12-9-r94.
12. Fujita PA, Rhead B, Zweig AS, et al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research* 2010;.
13. Schwarz JM, Rodelsperger C, Schuelke M, et al. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;7: 575–6.
14. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–9.
15. Olatubosun A, Valiaho J, Harkonen J, et al. PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat* 2012;.
16. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011;39:D945–50.
17. Maqungo M, Kaur M, Kwofie SK, et al. DDPC: dragon database of genes associated with prostate cancer. *Nucleic Acids Research* 2010;.
18. Cerami EG, Gross BE, Demir E, et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;39:D685–90.
19. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28: 27–30.
20. Pico AR, Kelder T, van Iersel MP, et al. WikiPathways: pathway editing for the people. *PLoS Biol* 2008;6:e184.
21. Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. *Nucleic Acids Res* 2013;41:D48–55.
22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–75.
23. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489: 57–74.
24. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;22:1790–7.
25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;25:1105–11.
26. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010; 11:R106,2010–11-10-r106.
27. Eeles RA, Olama AA, Benlloch S, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* 2013;45:385, 91, 391e1–2.
28. Zeller T, Wild P, Szymczak S, et al. Genetics and beyond—The transcriptome of human monocytes and disease susceptibility. *PLoS One* 2010;5: e10693.
29. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet* 2007;39:1217–24.
30. Barbieri CE, Baca SC, Lawrence MS, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* 2012;44:685–9.
31. Barbieri CE, Bangma CH, Bjartell A, et al. The mutational landscape of prostate cancer. *Eur Urol* 2013;64:567–76.
32. Kumar R, Manning J, Spendlove HE, et al. ZNF652, a novel zinc finger protein, interacts with the putative breast tumor suppressor CBFA2T3 to repress transcription. *Mol Cancer Res* 2006;4:655–65.
33. Callen DF, Ricciardelli C, Butler M, et al. Co-expression of the androgen receptor and the transcription factor ZNF652 is related to prostate cancer outcome. *Oncol Rep* 2010;23:1045–52.
34. Kumar R, Selth LA, Schulz RB, et al. Genome-wide mapping of ZNF652 promoter binding sites in breast cancer cells. *J Cell Biochem* 2011;112: 2742–7.
35. Haiman CA, Chen GK, Blot WJ, et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat Genet* 2011;43:570–3.
36. Halkidou K, Cook S, Leung HY, et al. Nuclear accumulation of histone deacetylase 4 (HDAC4) coincides with the loss of androgen sensitivity in hormone refractory cancer of the prostate. *Eur Urol* 2004;45:382, 9; author reply 389.

37. Yang Y, Tse AK, Li P, et al. Inhibition of androgen receptor activity by histone deacetylase 4 through receptor SUMOylation. *Oncogene* 2011;30:2207–18.
38. Chen J, Zhu S, Jiang N, et al. HoxB3 promotes prostate cancer cell progression by transactivating CDCA3. *Cancer Lett* 2013;330:217–24.
39. Frietze S, Lan X, Jin VX, et al. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J Biol Chem* 2010;285:1393–403.
40. Huang CN, Huang SP, Pao JB, et al. Genetic polymorphisms in androgen receptor-binding sites predict survival in prostate cancer patients receiving androgen-deprivation therapy. *Ann Oncol* 2012;23:707–13.
41. Biswas AK, Johnson DG. Transcriptional and nontranscriptional functions of E2F1 in response to DNA damage. *Cancer Res* 2012;72:13–7.
42. Gulzar ZG, McKenney JK, Brooks JD. Increased expression of NuSAP in recurrent prostate cancer is mediated by E2F1. *Oncogene* 2013;32: 70–7.
43. Nie Z, Stanley KT, Stauffer S, et al. AGAP1, an endosome-associated, phosphoinositide-dependent ADP-ribosylation factor GTPase-activating protein that affects actin cytoskeleton. *J Biol Chem* 2002;277:48965–75.
44. Mihara H, Kurihara T, Watanabe T, et al. cDNA cloning, purification, and characterization of mouse liver selenocysteine lyase. Candidate for selenium delivery protein in selenoprotein synthesis. *J Biol Chem* 2000;275: 6195–200.
45. Bouwmeester T, Bauch A, Ruffner H, et al. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* 2004;6:97–105.
46. Brandt U. Energy converting NADH:quinone oxidoreductase (complex I). *Annu Rev Biochem* 2006;75:69–92.
47. Caldovic L, Morizono H, Gracia Panglao M, et al. Cloning and expression of the human N-acetylglutamate synthase gene. *Biochem Biophys Res Commun* 2002;299:581–6.
48. Chen JL, Li J, Kiriluk KJ, et al. Deregulation of a Hox protein regulatory network spanning prostate cancer initiation and progression. *Clin Cancer Res* 2012;18:4291–302.
49. Kim MS, Je EM, Oh JE, et al. Mutational and expressional analyses of SPOP, a candidate tumor suppressor gene, in prostate, gastric and colorectal cancers. *APMIS* 2013;121:626–33.

Germline Copy Number Variation Analysis in Finnish Families With Hereditary Prostate Cancer

Virpi H. Laitinen,[1] Oyediran Akinrinade,[1,2] Tommi Rantapero,[1] Teuvo L.J. Tammela,[3] Tiina Wahlfors,[1] and Johanna Schleutker[1,4,5]*

[1]*BioMediTech, University of Tampere and Fimlab Laboratories, Tampere, Finland*
[2]*Children's Hospital, Institute of Clinical Medicine, University of Helsinki and Helsinki University Central Hospital, HUS, Finland*
[3]*Department of Urology, Tampere University Hospital and School of Health Sciences, University of Tampere, Tampere, Finland*
[4]*Medical Biochemistry and Genetics, Institute of Biomedicine, Turku, Finland*
[5]*Tyks Microbiology and Genetics, Department of Medical Genetics, Turku University Hospital, Turku, Finland*

BACKGROUND. The inherited factors that predispose individuals to prostate cancer (PrCa) remain largely unknown. The aim of this study was to identify germline copy number variants (CNVs) in Finnish individuals that could contribute to an increased PrCa risk.

METHODS. Genome-wide CNV screening was performed by analyzing single nucleotide polymorphisms from 105 PrCa patients and 37 unaffected relatives, representing 31 Finnish hereditary PrCa (HPC) families. The CNVs that aggregated in affected individuals and overlapped with genes implicated in cancer were validated using quantitative PCR in 189 index patients from Finnish HPC families and in 476 controls.

RESULTS. An intronic deletion (14.7 kb) in the *EPHA3* gene coding for class A ephrin receptor was observed in 11.6% of PrCa patients and in 6.1% of controls. The deletion associated with an increased PrCa risk ($P = 0.018$, $OR = 2.06$, $95\%CI = 1.18–3.61$). Although incomplete segregation with affection status was observed, the results show that the deletion was overrepresented in PrCa patients (56.1%) when compared to unaffected male relatives (31.2%). Interestingly, PrCa-specific mortality was higher among *EPHA3* deletion carriers (24.3%) than among patients with a normal *EPHA3* copy number (3.4%).

CONCLUSIONS. This study is the first investigation of the contribution of germline CNVs to HPC susceptibility in Finland. A novel association between the *EPHA3* deletion and PrCa risk was observed and, if confirmed, screening for this variant may aid in risk stratification among HPC patients. *Prostate 76:316–324, 2016.* © 2015 Wiley Periodicals, Inc.

KEY WORDS: **prostate cancer risk; copy number variation; germline; genetic predisposition; *EPHA3***

INTRODUCTION

Prostate cancer (PrCa), the most common male malignancy and the second leading cause of cancer death, was diagnosed in approximately 5,000 Finnish men in 2013 (Finnish Cancer Registry, Cancer Statistics, http://www.cancerregistry.fi/). The three major risk factors of PrCa include advanced age (>65 years), ethnic background and positive family history [1]. Contrary to other common cancers, genetic factors have a pronounced role in PrCa susceptibility. In a

Grant sponsor: Finnish Cancer Organizations; Grant sponsor: The Academy of Finland; Grant number: 251074; Grant sponsor: The Sigrid Juselius Foundation; Grant sponsor: The Ida Montin Foundation; Grant sponsor: The Competitive Research Funding of the Tampere University Hospital; Grant numbers: X51003; 9P053.

Conflicts of interest: The authors declare no conflict of interest.

*Correspondence to: Johanna Schleutker, Medical Biochemistry and Genetics, Institute of Biomedicine, Kiinamyllynkatu 10, FI-20014 University of Turku, Finland. E-mail: Johanna.Schleutker@utu.fi

Received 26 August 2015; Accepted 21 October 2015
DOI 10.1002/pros.23123
Published online 10 November 2015 in Wiley Online Library
(wileyonlinelibrary.com).

recent twin study, the heritability of PrCa was estimated to be as high as 58% [2].

Despite the large number of family-based studies and case-control association analyses, only a handful of PrCa susceptibility genes have been discovered. In Finnish hereditary PrCa (HPC) families, two genetic variants predominate. The carrier frequencies for the *CHEK2* variant p.I157T and the *HOXB13* variant p.G84E have been estimated to be as high as 10.8% and 8.4%, respectively [3,4]. The *RNASEL* variant p.E265X has been detected in 4.3% of HPC patients [5] and the truncating *CHEK2* mutation 1100delC, in 3.3% of HPC patients [3]. Variants in other susceptibility genes, including *ELAC2*, *MSR1*, *BRCA2*, and *PALB2* are rare and have only a limited role in PrCa predisposition in Finland [6–9]. Recently, genome-wide association studies have led to the identification of more than 100 single nucleotide polymorphisms (SNPs) that are linked to an increased PrCa risk (reviewed in ref. [10]). These SNPs account for approximately one third of the inherited risk. Nevertheless, in the majority of the Finnish HPC families the underlying genetic risk factors are still unknown.

Copy number variants (CNVs) have an established role in complex human diseases, including neurological disorders, asthma, type 2 diabetes and cancer [11]. Especially rare, biallelic CNVs ranging in size from 10 to >100 kb, and associated gene deletions and/or fusions may contribute to elevated cancer risk. Pathogenicity may be mediated via oncogene activation or the loss or inactivation of a tumour suppressor gene [12]. Recent studies have demonstrated an association between germline CNVs and an increased risk of cancers, such as neuroblastoma [13], colorectal cancer [14], and breast cancer [15,16]. The role of germline CNVs in predisposition to PrCa has also been explored. In Caucasian populations, deletions at 2p24.3 and 20p13 were associated with PrCa [17,18]. Subsequently, two additional risk loci at 15q21.3 and 12q21.31 have been identified [19]. Ledet et al. [20] described a germline duplication at 14q32.33, which was found to be enriched in PrCa patients from high-risk African American families.

The aim of this study was to uncover novel genetic risk factors contributing to PrCa predisposition in Finland. To explain missing heritability, we screened Finnish high-risk HPC families for germline CNVs and investigated the association of these CNVs with susceptibility to hereditary PrCa.

SUBJECTS AND METHODS

Hereditary Prostate Cancer Families

The collection of Finnish HPC families used in this study has been described previously [21]. When selecting the most representative families for the genome-wide SNP genotyping, families with a high rate of PrCa cases and with high availability of DNA samples were given priority. Thirty-one Finnish HPC families with 150 confirmed PrCa cases (three to eight patients per family) were included in the analysis. In total, 105 PrCa patients (median: three patients per family), 30 unaffected male relatives (healthy at the time of blood donation) and seven first-degree female relatives were genotyped. The clinical characteristics of the PrCa patients are presented in Supplementary Table SI. At the time of diagnosis, the average age of patients was 63.5 years, the average PSA value was 16.6 ng/ml and the median Gleason score for prostate biopsies was six.

The validation of select CNVs was performed in a larger sample set comprised of index patients from an additional 189 Finnish HPC families and 476 population-matched male control subjects. The majority of these HPC families (n = 149) had at least three family members affected with PrCa. In the remaining 40 families, there were only two affected family members, but at least one of the patients had been diagnosed with PrCa under the age of 60 years. The 476 population control samples had been collected from anonymous, unaffected male blood donors by the Finnish Red Cross Blood Services.

Of the 189 index patients included in the validation step, 22 patients harboured the *EPHA3* deletion at 3p11.1. The segregation of this CNV with affection status was studied in 21 of these HPC families. For one family, additional DNA samples were not available. Altogether, 210 family members were genotyped (range: 2–24 individuals per family), including 66 PrCa patients, 80 unaffected men, and 64 women.

All of the samples were collected with written and signed informed consent. The cancer diagnoses were confirmed using medical records and the annual update from the Finnish Cancer Registry. The project was approved by the local research ethics committee at Pirkanmaa Hospital District and by the National Supervisory Authority for Welfare and Health.

Copy Number Variation Analysis

The DNA samples from 105 PrCa cases and 37 unaffected family members were genotyped using the genome-wide SNP array HumanOmniExpress-12 BeadChip Kit (Illumina, Inc., San Diego, CA), which contained more than 700,000 optimized tag SNPs from all three HapMap phases. Sample preparation using the Infinium® Assay (Illumina, Inc.) and genotyping of the SNP markers were performed at the Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland according to the manufacturer's protocol. The CNV calling was performed as described in ref. [16], using the GenomeStudio v.2010.2 software and the PennCNV program [22]. The call rate for each sample was >99.5%, and thus, all of the samples were suitable for CNV calling. CNVs spanning less than three SNPs were excluded from the analysis.

The comparision of CNV distribution and median CNV lengths between PrCa patients and unaffected controls was performed using the Wilcoxon test (R v3.1.2, http://www.R-project.org/; ref. [23]). CNV carrier frequencies between patients and controls were compared with Fisher's exact test. In cases where a non-numerical *P*-value was obtained from Fisher's exact test, the odds ratios were estimated using the Visualizing Categorical Data (VCD) package [24] implemented in R. The 95% confidence intervals could not be reliably determined due to small number of control individuals carrying the CNVs.

Data Analysis

The CNVs identified in this study were compared to published CNV data stored in the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/) using GRCh37 (hg19) as the reference genome. A CNV was designated novel if less than 50% of its length overlapped with the previously reported CNVs in the DGV. To uncover genes located in the identified CNV loci, gene annotation was performed using the NCBI Reference Sequence Database (RefSeq, http://www.ncbi.nlm.nih.gov/refseq/). For intergenic CNVs, the nearest gene upstream or downstream of the CNV was determined using BEDTools [25]. All of the annotated genes were further queried for overlap against genes listed in the Online Mendelian Inheritance in Man® database (OMIM, http://www.omim.org/) to identify disease-associated genes.

To investigate the biological functions of the genes located in the identified CNV loci, an enrichment analysis was performed using the web-based Gene Set Analysis Toolkit V2 (WebGestalt2; ref. [26]). The applied categories included Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Pathway Commons and WikiPathways. The *P* values were adjusted using Benjamini–Hochberg method, and the threshold for significantly enriched category was set to 0.05.

Additionally, a family-based analysis was performed to evaluate the enrichment of CNVs in certain families. The percentage of CNVs in affected individuals was determined for each family by using the total number of PrCa patients in the family, the number of patients genotyped, and the number of patients harbouring the CNV. Enrichment was declared if at least 50% of the total number of patients in the family and/or patients genotyped carried the CNV.

CNV Validation and Familial Segregation Analysis by Real-Time Quantitative PCR

Four CNVs enriched in PrCa patient group were validated in an additional 189 index patients from Finnish HPC families and in 476 male controls by real-time quantitative PCR (qPCR). Similarly, the cosegregation of the *EPHA3* deletion with affection status was studied in 21 deletion-positive Finnish HPC families by real-time qPCR. Genotyping was performed on an ABI PRISM 7900HT sequencedetection system using the pre-designed TaqMan®Copy Number Assays Hs05836821_cn (2q34), Hs03480483_cn (3p11.1), Hs00434275_cn (5p13.3), and Hs03692888_cn (8p23.2) (Applied Biosystems/Life Technologies, Carlsbad, CA). The qPCR reactions were prepared in four replicates and run with a TaqMan® RNaseP Reference Assay (Applied Biosystems/Life Technologies), which was used as an internal standard. The method is described in detail elsewhere [16]. The qPCR results were analyzed with the CopyCaller™ Software v2.0 (Applied Biosystems/Life Technologies).

The Hardy–Weinberg Equilibrium (HWE) and case-control association tests for the four validated CNVs were performed using PLINK [27]. The *P*-value threshold for the HWE test was set to 0.05. The statistical significance of the association was evaluated with a two-sided Fisher's exact test.

RESULTS

The SNP array-based genome-wide CNV analysis targeting cytogenetically important regions was performed on 105 PrCa patients from Finnish high-risk HPC families and on 37 of their unaffected relatives. The PennCNV program detected a total of 2,575 autosomal CNVs at 544 different genomic regions in the sample set ($n = 142$). All the identified CNVs are listed in Supplementary Table SII. Data analysis revealed that deletions were more common than duplications. Altogether, 1,854 deletions and 721 duplications were detected, representing 72.0% and 28.0% of the CNVs, respectively. A majority of the CNVs (94.6%) were heterozygous, whereas only 139 deletions and a single duplication were homozygous. A summary of the identified CNVs is shown in Table I.

On average, 18 CNVs (13 deletions and five duplications) were detected per individual sample (Table I). CNVs were slightly more frequent in the controls than in PrCa patients, but the difference was not statistically significant. Analysis of CNV size

TABLE I. A Summary of the Identified 2,575 Copy Number Variants (CNVs) in 105 Prostate Cancer Patients and 37 Unaffected Relatives

CNVs	Average no. per sample	Median size (bp)	Overlap with genes (%)	Novel CNVs (%)
All (n = 2575)	18.1 (2575/142)		1228/2575 (47.7)	46/2575 (1.78)
PrCa patients	17.6 (1846/105)	11460	870/1846 (47.1)	29/1846 (1.57)
Controls	19.7 (729/37)	11360	358/729 (49.1)	17/729 (2.33)
Homozygous del (n = 139)	0.98 (139/142)		20/139 (14.4)	-
PrCa patients	0.92 (97/105)	3813	13/97 (13.4)	-
Controls	1.14 (42/37)	4770	7/42 (16.7)	-
Heterozygous del (n = 1715)	12.1 (1715/142)		783/1715 (45.7)	24/1715 (1.40)
PrCa patients	11.6 (1223/105)	9029	550/1223 (45.0)	17/1223 (1.39)
Controls	13.3 (492/37)	9168	233/492 (47.4)	7/492 (1.42)
Heterozygous dupl (n = 720)	5.1 (720/142)		424/720 (58.9)	22/720 (3.06)
PrCa patients	5.0 (525/105)	23720	306/525 (58.3)	12/525 (2.29)
Controls	5.3 (195/37)	20240	118/195 (60.5)	10/195 (5.13)
Homozygous dupl (n = 1)	0.007 (1/142)		1/1 (100)	-
PrCa patients	0.01 (1/105)	197024	1/1 (100)	-
Controls	-	-	-	-

CNVs were defined as novel if less than 50% of their length overlapped with previously reported CNVs in the Database of Genomic Variants.
PrCa, prostate cancer; Del, deletion; Dupl, duplication.

revealed that, in general, deletions were shorter than duplications (Table I). Again, the differences in the CNV size distribution between the two groups were not statistically significant. The median length of homozygous deletions was 3.8 kb in patients and 4.8 kb in controls ($P = 0.826$), and heterozygous deletions spanned approximately 9.0 kb in patients and 9.1 kb in controls ($P = 0.708$). The median size of heterozygous duplications was 23.7 kb in patients and 20.2 kb in controls ($P = 0.475$).

Annotation of the 2,575 CNVs against the NCBI RefSeq Database resulted in the identification of 1,228 gene-overlapping CNVs, of which 803 were deletions and 425 were duplications (Table I). Approximately half of the deletions (53.8%) and most of the duplications (88.2%) were exonic. A comparison of the 2,575 CNVs with the previously reported CNVs in DGV revealed 46 novel CNVs in our dataset. Of these, 36 overlapped with genes, including 19 deletions and 17 duplications. Novel heterozygous duplications were more than twice as frequent in unaffected controls (5.1%) than in PrCa patients (2.3%; Table I).

Validation of Selected CNVs

The family-based analysis revealed an enrichment of 63 CNVs in 26 families. These CNVs were located at 58 different genomic regions, and five of them were novel (Supplementary Table SIII). Because the aim of this study was to identify inherited copy number changes that predispose individuals to PrCa, we focused on CNVs clustered in families. A higher priority was given to CNVs that were enriched in affected individuals from more than one family. Furthermore, CNVs overlapping with genes that had either a known or potential biological role in PrCa susceptibility were favoured. Additional support for cancer-related gene function was obtained from gene ontology and pathway analyses. After careful evaluation, four CNVs were selected for validation by qPCR, including deletions affecting the *ERBB4* (2q34), *EPHA3* (3p11.1), and *CSMD1* (8p23.2) genes and a duplication overlapping the *PDZD2* (5p13.3) gene. The results from GO and pathway analyses for these four genes are shown in Supplementary Table SIV.

The four CNVs were genotyped in 665 individuals, including 189 index patients from HPC families and 476 male control samples. The genotyping results and carrier frequencies for each CNV are summarized in Table II. The frequency of homozygous *PDZD2* duplication carriers (3.2% in patients and 2.1% in controls) was unusually high when compared to the frequency of heterozygous carriers (2.6% in patients and 0.8% in controls). As expected, and different from the *ERBB4*, *EPHA3*, and *CSMD1* deletions, the *PDZD2* duplication was not in Hardy–Weinberg equilibrium either in PrCa patients or in controls. In the case-control association test, only one CNV, the 14.7 kb *EPHA3* deletion showed a statistically significant association with PrCa ($P = 0.018$, OR = 2.06, 95%CI = 1.18–3.61; Table III). The *EPHA3* deletion was detected in 22 PrCa patients

TABLE II. A Summary of the Genotyping Data and Carrier Frequencies for the Four Validated Copy Number Variants (CNVs) in Familial Index Cases (Affected; n = 189) and in Controls (Unaffected; n = 476)

CNV	Locus	Health status	DD_n (%)	DN_n (%)	NN_n (%)
ERBB4 deletion	2q34	Affected	-	4 (2.1)	185 (97.9)
		Unaffected	-	14 (2.9)	462 (97.1)
EPHA3 deletion	3p11.1	Affected	-	22 (11.6)	167 (88.4)
		Unaffected	-	29 (6.1)	447 (93.9)
PDZD2 duplication	5p13.3	Affected	6 (3.2)	5 (2.6)	178 (94.2)
		Unaffected	10 (2.1)	4 (0.8)	462 (97.1)
CSMD1 deletion	8p23.2	Affected	1 (0.5)	18 (9.5)	170 (90.0)
		Unaffected	-	49 (10.3)	427 (89.7)

DD, homozygous deletion/duplication; DN, heterozygous deletion/duplication; NN, normal copy number.

(11.6%) and in 29 controls (6.1%), and all of the *EPHA3* deletion carriers were heterozygous (Table II). The *ERBB4* and *CSMD1* deletions were not associated with the disease. The *ERBB4* deletion was more common in controls than in cancer patients ($P = 0.793$, OR = 0.72, 95%CI = 0.23–2.19), and the 2.7 kb *CSMD1* deletion had an equal frequency in both groups (5.3% in PrCa patients vs. 5.1% in controls; $P = 0.892$, OR = 1.03, 95% CI = 0.60–1.76; Table III).

Co-Segregation of *EPHA3* Deletion With Affection Status

To study the co-segregation of the 14.7 kb *EPHA3* deletion with affection status, additional family members from 21 HPC families whose index cases carried the deletion were genotyped using qPCR. In total, 89 individuals out of the 210 individuals genotyped were observed to carry the *EPHA3* deletion. The co-segregation of the deletion with affection status was incomplete in all of the 21 analyzed families. An example of a family pedigree is shown in Figure 1. However, when pooled together, 56.1% (37/66) of PrCa patients and only 36.1% (52/144) of unaffected family members carried the deletion. Twelve of the 144 unaffected family members had a diagnosis of another cancer type (predominantly breast or skin cancer), and six (50%) were deletion carriers. Three homozygous deletion carriers were observed in two families. Only one homozygous carrier was affected with PrCa, and the clinical course of his disease was indolent.

The clinical features of the 37 *EPHA3* deletion carriers were compared to those of the 29 PrCa patients with normal *EPHA3* copy number. The average age at diagnosis was essentially the same for both patient groups (63.8 years for carriers vs. 65.5 years for non-carriers), as was the Gleason score (7 for carriers vs. 6.5 for non-carriers). However, the average PSA value at diagnosis was mildly

TABLE III. Case-Control Association Test Results for the Four Validated Copy Number Variants and Prostate Cancer Risk

Cytoband[a]	Gene symbol/Entrez gene ID	CNV type	Size (kb)[b]	F_case[c]	F_control[d]	*P* value	OR (95%CI)
2q34	*ERBB4*/2066	Intronic deletion	25.6–55.7	0.011	0.015	0.793	0.72 (0.23–2.19)
3p11.1	*EPHA3*/2042	Intronic deletion	14.7	0.061	0.030	**0.018**	2.06 (1.18–3.61)
5p13.3	*PDZD2*/23037	Exonic duplication	52.1	0.045	0.025	0.077[e]	1.82 (0.97–3.43)
8p23.2	*CSMD1*/64478	Intronic deletion	2.7	0.053	0.051	0.892	1.03 (0.60–1.76)

The statistically significant *P*-value ($P < 0.05$) is shown in bold. All of these CNVs have been previously reported in the Database of Genomic Variants.

CNV, copy number variant; OR, odds ratio; CI, confidence interval.

[a] According to GRCh37 (hg19). Exact genomic coordinates are provided in Table S2.

[b] Size reported for prostate cancer patients analyzed with the SNP array. CNV size may vary between individuals.

[c] The frequency of the CNV (deletion/duplication) allele in prostate cancer patients.

[d] The frequency of the CNV (deletion/duplication) allele in unaffected male control subjects.

[e] Not in HWE.

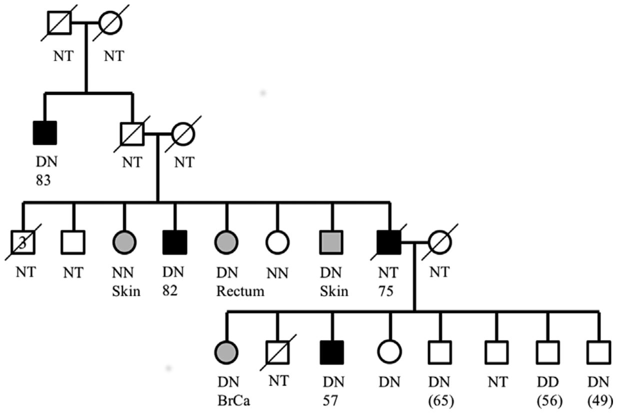


Fig. 1. Pedigree of Family ID 169. Incomplete co-segregation of the intronic *EPHA3* deletion with affection status was observed in all of the 21 families analyzed in this study. As an example, results for Family ID 169 are shown in detail. Squares denote males, and circles denote females. Deceased individuals are marked with a slash. Black squares indicate males with prostate cancer, and the age at diagnosis is marked under the square. Grey symbols indicate other cancers (BrCa = breast cancer). Genotypes are marked as follows: NN = normal copy number, DN = heterozygous deletion, DD = homozygous deletion, NT = not typed. The current age of unaffected male deletion carriers is given in parentheses.

elevated in carriers (43.2 ng/ml vs. 33.3 ng/ml in non-carriers). The most interesting clinical finding was the cause of death. Overall, 20 of the 66 PrCa patients died during the follow-up time, which varied from 17 to 22 years. Of the 37 *EPHA3* deletion carriers, nine patients (24.3%) died of PrCa, but among the 29 patients with normal *EPHA3* copy number, only one PrCa specific death (3.4%) had been reported. Secondary cancers were observed in 10.8% of deletion carriers and in 17.2% of patients with normal *EPHA3* copy number, but none of the patients who died of PrCa had been diagnosed with a secondary cancer.

The biological and molecular functions of the genes that overlapped with identified CNVs were explored by enrichment analysis. *EPHA3* and *ERBB4* were significantly overrepresented ($P < 0.05$) in several GO categories involving molecular functions related to receptor and signal transduction activities (for details, see Supplementary Table SIV). In addition, the cellular component category showed significant enrichment of *EPHA3*, *ERBB4*, and *CSMD1* in the plasma membrane, suggesting that these proteins may be involved in cell–cell interactions. Evidence for a role in the cell adhesion process was obtained for *EPHA3* and *PDZD2*. The KEGG pathway and the Pathway Commons analyses did not reveal any enriched categories with statistical significance for these four genes.

DISCUSSION

This study focused on identifying copy number variants that may explain at least a proportion of increased PrCa risk in Finnish HPC families. The genome-wide CNV profiling resulted in a total of 2,575 autosomal CNVs overlapping 544 unique loci. By using family-based enrichment analysis, we reduced the number of potentially pathogenic CNVs to 63. Subsequent data analysis steps focused on the identification of CNVs that predominantly clustered in affected individuals from multiple families and of affecting genes that could be linked to cancer-related pathways. The CNVs that were validated in a larger sample set included three deletions overlapping the intronic regions of the *EPHA3*, *CSMD1*, and *ERBB4* genes and a duplication overlapping exon 24 of the *PDZD2* gene. Although none of these CNVs was novel, each was detected in more than one family, and the affected genes were either known or likely to be involved in prostate carcinogenesis.

The CNV validation analysis revealed a statistically significant association between PrCa risk and the 14.7 kb deletion at intron five of the *EPHA3* gene (Table III). *EPHA3 (EPH Receptor A3)* gene is a member of the protein-tyrosine kinase family and encodes a class A ephrin receptor. *EPHA3* functions as a signal transduction molecule that participates in controlling adhesion, movement, shape, and growth of cells. Somatic mutations of *EPHA3* are frequently found in various carcinomas, including melanoma, glioblastoma, lung, colorectal, and hepatocellular cancers [28]. In a recent study, *EPHA3* was shown to contribute to the development and malignant progression of PrCa, possibly by activating the Akt pathway and thus blocking apoptosis [29].

We detected a heterozygous *EPHA3* deletion in 11.6% of PrCa patients and in 6.1% of controls (Table II). Familial segregation analysis revealed that more than half of the PrCa patients (56.1%) were deletion carriers, whereas only one third of unaffected family members (36.1%) carried the deletion. The proportion of unaffected male carriers was even lower, only 31.2% (25/80). Although complete segregation with affection status could not be demonstrated for any of the 21 families analyzed, the results show that *EPHA3* deletion aggregates in affected individuals. Of particular interest was the observation that PrCa-specific mortality was substantially higher among *EPHA3* deletion carriers than among patients with a normal *EPHA3* copy number. This finding, if confirmed by replication studies in larger patient cohorts, implicates *EPHA3* as having an important role in advanced stages of the disease.

A similar enrichment of the same *EPHA3* deletion was previously observed in Finnish patients with hereditary breast and/or ovarian cancer [16], but statistical significance for the association was not obtained. Nevertheless, our combined findings suggest that disruption of the genomic *EPHA3* sequence has an effect on *EPHA3* protein function. It is possible that, as argued in ref. [16], the deletion abolishes an intronic regulatory element, thereby leading to aberrant receptor activity. Both tumour suppressor and tumour promoting properties have been suggested for *EPHA3* [28]. In colorectal cancer patients, increased *EPHA3* expression has been associated with poorer survival [30].

The only exonic CNV included in the validation step was the 52.1 kb duplication at exon 24 of the *PDZD2 (PDZ Domain Containing 2)* gene. The *PDZD2* protein is located in the endoplasmic reticulum and may participate in intracellular signalling. High expression levels of *PDZD2* have been reported in prostate tumour cell lines and human primary prostate tumours, implicating an important role in the early stages of prostate tumourigenesis [31]. On the other hand, tumour suppressor function has also been suggested for *PDZD2* [32].

Validation results showed that the frequency of *PDZD2* duplication carriers was twice as high among PrCa patients (5.8%) than among unaffected controls (2.9%; Table II). However, a majority of duplication carriers were homozygous for the variant, and therefore it was not surprising to learn that this CNV was not in HWE. It is possible that the discrepancy from HWE is due to a genotyping error. However, this is unlikely as each sample was assayed in four replicates. Another explanation may be that the duplication is causative and therefore under selection. As such, we genotyped 97 individuals from the 12 PrCa families whose index patients carried the *PDZD2* duplication. Review of the genotyping data revealed that 64.5% of the affected individuals (20/31) and 33.3% of the unaffected family members (22/66) were duplication carriers. Homozygous duplications were observed in 45% (9/20) and 64% (14/22) of affected and unaffected carriers, respectively. In summary, the *PDZD2* duplication was detected to cluster among PrCa patients, but caution has to be taken in the interpretation of this observation because of the missing HWE. It will, however, be exciting to see whether future studies confirm the suggestive association between *PDZD2* and PrCa risk reported here.

Although the 2.7 kb deletion at intron five of the *CSMD1 (CUB And Sushi Multiple Domains 1)* gene was outside the most pathogenic CNV size range (from 10 to 100 kb; ref. [12]), we validated this CNV because of the intriguing properties of the affected gene. *CSMD1*, a potential tumour suppressor gene, is located at 8p23, a region frequently deleted in prostate tumours [33]. It encodes a transmembrane protein whose expression is lost especially in epithelial cancers. In addition, reduced expression of CSMD1 correlates with shortened survival in breast cancer [34] and with earlier onset of colorectal cancer [35]. Unfortunately, we were unable to show any difference in the frequency of *CSMD1* deletion carriers between PrCa patients and controls (Table II). The odds ratio of 1.03 further indicated that the PrCa risk was not elevated among deletion carriers (Table III). Hence, the 2.7 kb *CSMD1* deletion most likely represents a common polymorphism.

Similar to *CSMD1*, *ERBB4 (V-Erb-B2 Avian Erythroblastic Leukaemia Viral Oncogene Homologue 4)* is a promising PrCa candidate gene. *ERBB4* belongs to the protein-tyrosine kinase family and codes for a cell surface receptor protein which is activated by neuregulins and epidermal growth factors. Activation of the ERBB4 receptor induces several cellular processes, such as cell growth, proliferation, and differentiation. Somatic mutations in the *ERBB4* gene have been shown to associate with gastric, colorectal, breast, and non-small cell lung cancers [36]. In association with the *HNF1b* gene, *ERBB4* has also been linked to increased PrCa risk [37]. Recent data suggest that *ERBB4* may act as a tumour suppressor [38]. We observed deletions ranging from 25.6 kb to 55.7 kb at intron 20 of the *ERBB4* gene in 2.1% of the PrCa patients (Tables II and III). However, these deletions were more frequent in unaffected controls (2.9%; Table II). Therefore, regardless of the significant enrichment of *ERBB4* in cell membrane receptor and signal transduction activities (Supplementary Table SIV), the association of the deletions identified in this study and PrCa risk could not be proven.

Like other genetic variants, CNVs may also be population specific. Different CNVs likely predominate in different populations. The Finnish population is a well-known genetic isolate [39] and, therefore, it is not surprising that CNVs that are rare elsewhere show significant enrichment in Finnish HPC families. Although the population-specificity of CNV distributions may complicate replication studies, it should be noted that unexpected findings observed in genetically isolated populations may aid in the identification of novel PrCa-associated molecules and provide fresh insights into the function of complex protein networks and PrCa-associated metabolic pathways.

In conclusion, this study complements our previous efforts on elucidating diverse genetic factors contributing to PrCa predisposition in Finland. This study is the first report on genome-wide, germline copy number profiling of Finnish PrCa families.

Novel associations between CNVs and PrCa were observed, and strongly suggestive evidence for the involvement of *EPHA3* in increased PrCa risk was obtained. Further independent and, preferably functional studies will be needed to confirm our preliminary findings. However, the *EPHA3* deletion may be considered a valid candidate for targeted PrCa screening panel intended for risk assessment in the Finnish population.

ACKNOWLEDGMENTS

The authors thank all of the patients and their family members for participating in this study. Ms. Kirsi Määttä and Mr. Daniel Fischer are acknowledged for helping with the data analysis. Ms. Riitta Vaalavuo and Ms. Riina Kylätie are thanked for technical assistance. The genotyping of SNP markers was performed by the Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland. This work was financially supported by the Finnish Cancer Organizations; The Academy of Finland (251074); The Sigrid Juselius Foundation; the Ida Montin Foundation; and the Competitive Research Funding of the Tampere University Hospital (X51003, 9P053).

REFERENCES

1. Bostwick DG, Burke HB, Djakiew D, Euling S, Ho SM, Landolph J, Morrison H, Sonawane B, Shifflett T, Waters DJ, Timms B. Human prostate cancer risk factors. Cancer 2004;101: 2371–2490.
2. Hjelmborg JB, Scheike T, Holst K, Skytthe A, Penney KL, Graff RE, Pukkala E, Christensen K, Adami HO, Holm NV, Nuttall E, Hansen S, Hartman M, Czene K, Harris JR, Kaprio J, Mucci LA. The heritability of prostate cancer in the Nordic Twin Study of Cancer. Cancer Epidemiol Biomarkers Prev 2014;23:2303–2310.
3. Seppala EH, Ikonen T, Mononen N, Autio V, Rokman A, Matikainen MP, Tammela TL, Schleutker J. CHEK2 variants associate with hereditary prostate cancer. Br JCancer 2003; 89:1966–1970.
4. Laitinen VH, Wahlfors T, Saaristo L, Rantapero T, Pelttari LM, Kilpivaara O, Laasanen SL, Kallioniemi A, Nevanlinna H, Aaltonen L, Vessella RL, Auvinen A, Visakorpi T, Tammela TL, Schleutker J. HOXB13 G84E mutation in Finland: Population-based analysis of prostate, breast, and colorectal cancer risk. Cancer Epidemiol Biomarkers Prev 2013;22:452–460.
5. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, Moses T, Ewing C, Gillanders E, Hu P, Bujnovszky P, Makalowska I, Baffoe-Bonnie A, Faith D, Smith J, Stephan D, Wiley K, Brownstein M, Gildea D, Kelly B, Jenkins R, Hostetter G, Matikainen M, Schleutker J, Klinger K, Connors T, Xiang Y, Wang Z, De Marzo A, Papadopoulos N, Kallioniemi OP, Burk R, Meyers D, Gronberg H, Meltzer P, Silverman R, Bailey-Wilson J, Walsh P, Isaacs W, Trent J. Germline mutations in the ribonuclease L gene in families showing linkage with HPCI. Nat Genet 2002;30:181–184.
6. Rokman A, Ikonen T, Mononen N, Autio V, Matikainen MP, Koivisto PA, Tammela TL, Kallioniemi OP, Schleutker J. ELAC2/HPC2 involvement in hereditary and sporadic prostate cancer. Cancer Res 2001;61:6038–6041.
7. Seppala EH, Ikonen T, Autio V, Rokman A, Mononen N, Matikainen MP, Tammela TL, Schleutker J. Germ-line alterations in MSR1 gene and prostate cancer risk. Clin Cancer Res 2003;9:5252–5256.
8. Ikonen T, Matikainen MP, Syrjakoski K, Mononen N, Koivisto PA, Rokman A, Seppala EH, Kallioniemi OP, Tammela TL, Schleutker J. BRCA1 and BRCA2 mutations have no major role in predisposition to prostate cancer in Finland. J Med Genet 2003;40:e98.
9. Pakkanen S, Wahlfors T, Siltanen S, Patrikainen M, Matikainen MP, Tammela TL, Schleutker J. PALB2 variants in hereditary and unselected Finnish prostate cancer cases. J Negat Results Biomed 2009;8:12.
10. Demichelis F, Stanford JL. Genetic predisposition to prostate cancer: update and future perspectives. Urol Oncol 2015;33:75–84.
11. Almal SH, Padh H. Implications of gene copy-number variation in health and diseases. J Hum Genet 2012;57:6–13.
12. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts Van Kessel A. Germline copy number variation and cancer risk. Curr Opin Genet Dev 2010;20:282–289.
13. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AI, London WB, Shaikh TH, Bradfield J, Grant SF, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM. Copy number variation at 1q21.1 associated with neuroblastoma. Nature 2009;459:987–991.
14. Venkatachalam R, Verwiel ET, Kamping EJ, Hoenselaar E, Gorgens H, Schackert HK, Van Krieken JH, Ligtenberg MJ, Hoogerbrugge N, Van Kessel AG, Kuiper RP. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. Int J Cancer 2011;129: 1635–1642.
15. Krepischi AC, Achatz MI, Santos EM, Costa SS, Lisboa BC, Brentani H, Santos TM, Goncalves A, Nobrega AF, Pearson PL, Vianna-Morgante AM, Carraro DM, Brentani RR, Rosenberg C. Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Res 2012;14:R24.
16. Kuusisto KM, Akinrinade O, Vihinen M, Kankuri-Tammilehto M, Laasanen SL, Schleutker J. Copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. PloS ONE 2013;8:e71802.
17. Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Sun J, Wiklund F, Wiley K, Isaacs SD, Stattin P, Xu J, Duggan D, Carpten JD, Isaacs WB, Gronberg H, Zheng SL, Chang BL. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. Cancer Res 2009;69:2176–2179.
18. Jin G, Sun J, Liu W, Zhang Z, Chu LW, Kim ST, Sun J, Feng J, Duggan D, Carpten JD, Wiklund F, Gronberg H, Isaacs WB, Zheng SL, Xu J. Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. Carcinogenesis 2011;32:1057–1062.
19. Demichelis F, Setlur SR, Banerjee S, Chakravarty D, Chen JY, Chen CX, Huang J, Beltran H, Oldridge DA, Kitabayashi N, Stenzel B, Schaefer G, Horninger W, Bektic J, Chinnaiyan AM, Goldenberg S, Siddiqui J, Regan MM, Kearney M, Soong TD,

Rickman DS, Elemento O, Wei JT, Scherr DS, Sanda MA, Bartsch G, Lee C, Klocker H, Rubin MA. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci USA 2012;109:6686–6691.

20. Ledet EM, Hu X, Sartor O, Rayford W, Li M, Mandal D. Characterization of germline copy number variation in high-risk African American families with prostate cancer. Prostate 2013;73:614–623.

21. Schleutker J, Matikainen M, Smith J, Koivisto P, Baffoe-Bonnie A, Kainu T, Gillanders E, Sankila R, Pukkala E, Carpten J, Stephan D, Tammela T, Brownstein M, Bailey-Wilson J, Trent J, Kallioniemi OP. A genetic epidemiological study of hereditary prostate cancer (HPC) in Finland: frequent HPCX linkage in families with late-onset disease. Clin Cancer Res 2000;6:4810–4815.

22. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007;17:1665–1674.

23. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2014.

24. Meyer D, Zeileis A, Hornik K. VCD:Visualizing Categorical Data. R package version 1.4-1;2015.

25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–842.

26. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33:W741–W748.

27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J HumGenet 2007; 81:559–575.

28. Lisabeth EM, Fernandez C, Pasquale EB. Cancer somatic mutations disrupt functions of the EphA3 receptor tyrosine kinase through multiple mechanisms. Biochemistry 2012;51:1464–1475.

29. Wu R, Wang H, Wang J, Wang P, Huang F, Xie B, Zhao Y, Li S, Zhou J. EphA3, induced by PC-1/PrLZ, contributes to the malignant progression of prostate cancer. Oncol Rep 2014;32:2657–2665.

30. Xi HQ, Zhao P. Clinicopathological significance and prognostic value of EphA3 and CD133 expression in colorectal carcinoma. J Clin Pathol 2011;64:498–503.

31. Chaib H, Rubin MA, Mucci NR, Li L, Taylor JMG, Day ML, Rhim JS, Macoska JA. Activated in prostate cancer: a PDZ domain-containing protein highly expressed in human primary prostate tumors. Cancer Res 2001;61:2390–2394.

32. Tam CW, Cheng AS, Ma RY, Yao KM, Shiu SY. Inhibition of prostate cancer cell growth by human secreted PDZ domain-containing protein 2, a potential autocrine prostate tumor suppressor. Endocrinology 2006;147:5023–5033.

33. Chang BL, Liu W, Sun J, Dimitrov L, Li T, Turner AR, Zheng SL, Isaacs WB, Xu J. Integration of somatic deletion analysis of prostate cancers and germline linkage analysis of prostate cancer families reveals two small consensus regions for prostate cancer genes at 8p. Cancer Res 2007;67:4098–4103.

34. Kamal M, Shaaban AM, Zhang L, Walker C, Gray S, Thakker N, Toomes C, Speirs V, Bell SM. Loss of CSMD1 expression is associated with high tumour grade and poor survival in invasive ductal breast carcinoma. Breast Cancer Res Treat 2010;121:555–563.

35. Shull AY, Clendenning ML, Ghoshal-Gupta S, Farrell CL, Vangapandu HV, Dudas L, Wilkerson BJ, Buckhaults PJ. Somatic mutations, allele loss, and DNA methylation of the Cub and Sushi Multiple Domains 1 (CSMD1) gene reveals association with early age of diagnosis in colorectal cancer patients. PloS ONE 2013;8:e58731.

36. Soung YH, Lee JW, Kim SY, Wang YP, Jo KH, Moon SW, Park WS, Nam SW, Lee JY, Yoo NJ, Lee SH. Somatic mutations of the ERBB4 kinase domain in human cancers. Int J Cancer 2006;118:1426–1429.

37. Hu YL, Zhong D, Pang F, Ning QY, Zhang YY, Li G, Wu JZ, Mo ZN. HNF1b is involved in prostate cancer risk via modulating androgenic hormone effects and coordination with other genes. Genet Mol Res 2013;12:1327–1335.

38. Gallo RM, Bryant IN, Mill CP, Kaverman S, Riese DJ. Multiple functional motifs are required for the tumor suppressor activity of a constitutively-active ErbB4 mutant. J Cancer Res Ther Oncol 2013;1:10.

39. Peltonen L, Jalanko A, Varilo T. Molecular genetics of the Finnish disease heritage. Hum Mol Genet 1999;8:1913–1923.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.