Tampere University

TOMMI RANTAPERO

Developing a Framework for Analysis of Next-Generation Sequencing Data in Cancer Genetics and Epigenetics

Tampere University Dissertations 311

Tampere University Dissertations 311

TOMMI RANTAPERO

Developing a Framework for Analysis of Next-Generation Sequencing Data in Cancer Genetics and Epigenetics

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the auditorium F115
of the Arvo building, Arvo Ylpön katu 34, Tampere,
on 16 October 2020, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine and Health Technology
University of Turku, Institute of Biomedicine
Finland
Karolinska Institutet, Department of Medical Epidemiology and Biostatistics
Sweden

Responsible supervisor and Custos	Professor Matti Nykter Tampere University Finland	
Pre-examiners	Associate Professor Sami Heikkinen University of Eastern Finland Finland	Associate Professor Antti Rannikko University of Helsinki Finland
Opponent	Docent Esa Pitkänen University of Helsinki Finland	

Cover design: Roihu Inc.

ISBN 978-952-03-1702-7 (print)
ISBN 978-952-03-1703-4 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1703-4

PunaMusta Oy – Yliopistopaino
Vantaa 2020

ACKNOWLEDGEMENTS

This doctoral thesis was carried out in the Computational biology group, faculty of Medicine and Health Technology at the Tampere University. First I would like to express my warmest gratitude to my supervisor Professor Matti Nykter and Professor Johanna Schleutker who granted me the opportunity to work in so many interesting projects. I thank you for your guidance and encouragement you have given to me. I feel privileged for having the change to work in both your research groups.

I would like to thank all my colleagues in the Computational biology and the Genetic Predisposition to Cancer group. I have learned a lot from you and I could always count on you whenever I needed help with something. I would especially want to thank my coauthors Virpi Laitinen, Kirsi Määttä, Daniel Fischer, Riikka Nurminen, Elisa Vuorela, Tiina Wahlfors from the Genetic Predisposition to Cancer group as well as Professor Teuvo Tammela. Moreover, I would like to thank my coauthors from the Cancer genomic group: Minna Ampuja, Alejandra Rodriguez-Martinez, Maaria Palmroth, and Professor Anne Kallioniemi. It was nice to work with you and I feel that our collaborative project was particularly educating for me. In addition, I would like to express my gratitude to our Swedish collaborators and coauthors at Karolinska Institutet. Especially I would like to thank Associate Professor Fredrik Wiklund for his great work regarding our manuscript.

I would like also like to thank the members of Experimental Immunology and Microbiology and Immunology groups who I had the chance to work with. I really enjoyed working with our common projects and the projects really broadened my understanding of the use of bioinformatics in sequencing data analysis.

Many thanks to members of my thesis committee members Professor Marko Pesu and Professor Olli Yli-Harja. Our meetings were very relaxed and your comments and encouraging words gave me hope that someday I will manage to finish my thesis.

I would also want to thank all my current coworkers at Genevia Technologies. Special thanks go to Jane Pulman for doing an excellent job reviewing the language of my thesis. I would also want to thank Klaus Breitholtz and Antti Ylipää for offering me plenty of flexibility and time to finish my thesis.

Finally I would like my friends and family. I am grateful for all the love and support I have received from you during these years.

ABSTRACT

The development of Next-generation sequencing technology has opened up new possibilities in the field of biomedical research. This novel technology has been widely applied in cancer research to study various aspects of this complex disease. However, efficient algorithms, statistical methods and various databases are needed to be able to harness the massive amounts of data being produced by this technology. Such computational methods are applied in bioinformatic tools, which in turn are integrated into analysis frameworks which can be used to answer various biological questions.

The first aim of this study was to develop a bioinformatics framework for analysis of Next-generation sequencing data in order to discover and characterise germline variants associated with hereditary cancer. This framework was applied and developed further in three studies. In the first study two loci 2q37 and 17q11.2-q22, which have been previously associated with prostate cancer, were sequenced and the variants were characterised by conducting an association study within in a larger set of individuals. In the second study individuals with breast cancer and/or ovarian cancer, which are not known to carry BRCA1/2 germline variants, were sequenced using Whole Exome sequencing in order to discover candidate genes associated with cancer susceptibility. In the third study, Finnish and Swedish individuals with lethal prostate cancer were sequenced using Whole Exome sequencing and compared against cases which were not deemed lethal based on the aggressiveness of the disease and population controls in order to uncover genes associated with the extremely aggressive form of the disease.

The second aim was to extend the established framework for integrating data from several NGS applications to uncover the role of both genetics and epigenetics in cancer development. This extended framework was applied in two studies. Firstly, the framework was applied in the first study to characterise the regulatory potential of the non-coding variants located in 2q37 and 17q11.2-q22. Secondly, the framework was applied to analyse and integrate RNA-seq and Dnase-seq data in order to study BMP4 response in two breast cancer cell lines. The overall aim of this study is to gain insight into how epigenetic factors and transcriptional regulators mediate the effects of BMP4 stimulus.

By utilising the developed framework low to moderate risk variants significantly associated with prostate cancer were discovered in HDAC4 and ZNF652. Moreover, the individuals with breast and/or ovarian cancer were found to have enriched number of pathogenic variants in ATM, MYC, PLAU, RAD1 and RRM2B suggesting that these genes may be associated with cancer susceptibility. Finally, the framework discovered variants likely to be associated with extremely aggressive prostate cancer and comparison of carrier rates of these variants revealed that among the Finnish and Swedish populations ATM and CHEK2 seemed to be strongly associated with extremely aggressive prostate cancer. Interestingly, in BRCA2 which has been shown to have the strongest association to aggressive prostate cancer in previous studies did not harbour likely pathogenic variants among the lethal cases.

The extended framework revealed non-coding variants which are associated to gene expression (eQTL variants) of which one targeted *TBKBP1* that was also shown to be differentially expressed between affected individuals and controls. Moreover, this variant has been reported as an eQTL by previous studies. Another putative eQTL variant was found to be associated with *ZNF652* which was also shown to be associated with prostate cancer based on coding variants harboured by the gene which had been observed in the same cohort. Moreover, the use of the extended framework in the integration of epigenetic and transcriptomic data revealed that BMP4 response genes are dependent on the epigenetic profile and that transcription factors *MBD2*, *CBFB* ja *HIF1A* have a role in the regulation of some these target genes. Furthermore, BMP4 stimulation was shown to cause varied responses in the epigenetic profiles of the different breast cancer cell lines which are consistent with findings related to the behaviour induced by the stimulation.

In conclusion, the framework developed for analysis of germline variant data identified novel candidate genes as well as variants associated with hereditary prostate, breast and ovarian cancer. The extended framework identified eQLTs which might be associated with the development of prostate cancer. Moreover, epigenetic alteration as well as transcription factors involved in cancer progression were characterised utilising the developed framework.

TIIVISTELMÄ

Uuden sukupolven sekvensointi menetelmien kehitys on tuonut mukanaan uusia mahdollisuuksia biolääketieteen tutkimusalalla. Kuluneen vuosikymmenen aikana tätä uutta teknologiaa on hyödynnetty laajasti syöpätutkimuksessa pyrkimyksenä paremmin ymmärtää taudin eri piirteitä. Jotta uusien sekvensointi menetelmien tuottaman valtavan datan perusteella voitaisiin tehdä biologisesti merkityksellisiä johtopäätöksiä, on analyysissä käytettävä bioinformatiikan menelmiä, jotka perustuvat tehokkaiden algoritmien, tilastotieteen sekä erilaisten tietokantojen hyödyntämiseen.

Tämä tutkimuksen ensimmäisenä tavoitteena oli hyödyntää bioinformatiikan menetelmiä uuden sukupolven sekvensointi datan analysointia varten tutkittaessa ituradan varianttien yhteyttä perinnöllisiin syöpiin. Ensimmäisessä tutkimuksessa eturauhassyöpään liitetyt kromosomaaliset lokukset 2q37 ja 17q11.2-q22 sekvensoitiin kohdennetusti perinnölliseen eturauhassyöpään sairastuneita yksilöiltä. Löydetyt variantit karakterisoitiin tekemällä assosiaatioanalyysi käyttämällä hyödyksi suurempia syöpä- sekä kontrollikohortteja. Toisessa tutkimuksessa perinnölliseen kolmoisnegatiiviseen rintasyöpään ja/tai munasarjasyöpään sairastuneita yksilöitä, karakterisoitiin koko eksomisekvesoinnilla pyrkimyksenä löytää uusia kandidaattigeenejä sekä variantteja, jotka altistavat perinnölliselle syövälle. Kolmannessa tutkimuksessa tarkoituksena oli löytää kandidaattigeenejä, jotka altistavat äärimmäisen aggressiiviselle eturauhassyövälle. Menetelmänä käytettiin koko eksomin sekvensointia. Tutkimuskohortin muodostivat potilaat, jotka olivat kuolleet eturauhassyöpään ja kontrollikohortin muodostivat eturauhassyöpäpotilaat, jotka eivät kuolleet eturauhassyöpään. Lisäksi käytettiin populaatiokontrollia.

Tutkimuksen toisena tavoitteena oli pyrkiä hyödyntämään bioinformatiikan menetelmiä uuden sukupolven sekvensointisovellusten tuottaman datan integratiivista analyysissä, geneettisten ja epigeneettisten tekijöiden roolien selvittämiseksi syövän kehityksessä. Näitä menetelmiä hyödynnettiin kahdessa tutkimuksessa, joista ensimmäisessä tutkittiin 2q37 and 17q11.2-q22 lokuksissa havaittujen ei-koodavien varianttien mahdollista osuutta eturauhassyöpään liittyvien geenien säätelyssä. Toisessa tutkimuksessa menetelmiä hyödynnettiin RNA-seq ja Dnase-seq datojen integratiivisessa analyysissä tarkoituksena tutkia BMP4 vastetta

kahdessa eri rintasyöpä-solulinjassa. Tutkimuksen tavoitteena oli selvittää, mitkä epigeneettiset tekijät ja transkription säätelijät välittävät BMP4:n vastetta soluissa.

Ensimmäisen tutkimuksen tuloksena löydettiin uusia perinnölliseen eturauhassyöpään liittyviä matalan ja keskisuuren riskin variantteja *HDAC4* ja *ZNF652* geeneistä. Yhdistämällä kohdennetussa sekvensoinnin sekä RNA-seq:n tuottama data, löydettiin eQTL variantteja, jotka mahdollisesti liittyvät geenien säätelyn eturauhassyövässä. Eräs varianteista näytti liittyvän *TBKBP1*:n säätelyyn, jonka ilmentymisessä havaittiin olevan eroja syöpään sairastuneiden ja kontrollien välillä. Viitteitä tämän variantin roolista geenin säätelyssä on myös löydetty aiemmassa tutkimuksessa. Toisen mahdollisen eQTL variantin havaittiin olevan yhteydessä *ZNF652* säätelyyn, josta oli myös löydetty eturauhassyöpään assosioituvia koodaavia variantteja samasta aineistosta.

Toisessa tutkimuksessa rintasyöpään ja/tai munasarjasyöpään sairastuneilla yksilöillä havaittiin, että patogeenisiksi oletetetut variantit olivat rikastuneet A*TM, MYC, PLAU, RAD1* ja *RRM2B* geeneihin. Tämä viittaa siihen, että jo tunnetun *BRCA2*:n lisäksi nämä kyseiset geenit liittyvät kasvaneeseen syöpä-alttiuteen.

Kolmannen tutkimuksessa tutkimuksessa epigeneettisen ja transkriptio-datan integroinnissa paljasti BMP4:n kohdegeenien olevan riippuvainen solujen epigeneettisestä profiilista sekä transkriptiotekijöitä: *MBD2*, *CBFB* ja *HIF1A*, jotka osallistuvat eräiden kohdegeenien ilmentymisen säätelyyn. Lisäksi BMP4 stimulaation havaittiin aiheuttavan hyvin vaihtelevia muutoksia solulinjojen epigeettisissä profiileissa, jotka ovat linjassa solujen käyttäytymisessä havaituissa eroissa niitä stimuloitaessa.

Neljännen tutkimuksen tuloksena eturauhassyöpään kuolleilta suomalaisia ja ruotsalaisia löydettiin äärimmäisen aggressiiviseen eturauhassyöpään liittyviä geenejä verrattaessa patogeenisiksi oletettujen varianttien määriä eturauhassyöpään kuolleiden sekä kontrollikohorttien välillä. Näistä geeneistä *CHEK2* ja *ATM* osoittautuivat liittyvän voimakkaimmin äärimmäisen aggressiivisen tautiin. Aiemmista tutkimuksista poiketen, äärimmäiseen aggressiiviseen eturauhassyöpään vahvimmin liitetyllä *BRCA2*:lla ei havaittu olevan merkittävää assosiaatiota syöpä-alttiuteen tutkittavissa populaatioissa.

Yhteenvetona bioinformatiikan menelmiä hyödyntämällä, löydettiin uusia kandidaattigeenejä sekä ituradan variantteja, jotka ovat yhteydessä perinnöllisiin eturauhas-, rinta- sekä munasarjasyöpiin. Lisäksi löydettiin geenin säätelyyn liittyviä eQTL variantteja, jotka ovat mahdollisesti yhteydessä eturauhassyövän kehitykseen ja karakterisoitiin syövän kasvuun ja kehitykseen liittyviä epigeneettisiä muutoksia sekä transkriptiotekijöitä.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

FN	False Negative
FP	False Positive
ACMG-AMP	American College of Medical Genetics and Genomics and Association of Molecular Pathology
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BQSR	Base Quality Score Recalibration
BS-seq	Bisulfite sequencing
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler Transform
CADD	Combined Annotation Dependent Depletion
CCD	Couple Charged Device Camera
cDNA	complementary Deoxyribonucleic Acid
COSMIC	Catalogue of Somatic Mutations in Cancer
CRT	Cyclic Reversible Termination
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDPC	Dragon database of Genes Implicated in Prostate Cancer
DHS	DNAseI hypersensitive site
DNA	Deoxyribonucleic Acid
DNase-seq	DNase sequencing
ECM	Extracellular Matrix
EM	Expectation maximization
EMT	Epithelial-to-Mesenchymal Transition
ENCODE	Encyclopedia of DNA elements (Intro)
eQTL	expression Qualitative Loci
ExAC	Exome Aggregation Consortium
FDR	False Discovery Rate
FFPE	Formalin-fixed paraffin-embedded
FGF	Fibroblast Growth Factors

FIMM	Institute for Molecular Medicine Finland
FM	Ferrari-Manzini
GATK	Genomic Analysis ToolKit
GnomAD	Genome Aggregation Database
GO	Gene Ontology
GREAT	Genomic Regions Enrichment of Annotations Tool
GWAS	Genome-Wide Association Study
HBOC	Hereditary Breast and Ovarian Cancer
HOCOMOCO	HOmo sapiens COmprehensive MOdel COllection
HPC	Hereditary Prostate Cancer
HWE	Hardy Weinberg Equilibrium
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
lincRNA	long non-coding Ribonucleic Acid
LOF	Loss Of Function
MAF	Minor Allele Frequency
MARA	Motif Activity Response Analysis
MDSCs	Myeloid-Derived Suppressor Cells
MeDIP-seq	Methylated DNA Immunoprecipitation Sequencing
miRNA	MicroRNA
MMP	Maximal Mappable Prefix
NGS	Next-Generation Sequencing
NMD	Nonsense Mediated Decay
OR	Odds Ratio
PARP	Poly ADP Ribose Polymerase
PCR	Polymerase Chain Reaction
PrCa	Prostate cancer
PTV	Protein Truncating Variant
PWM	Position Weight Matrix
QC	Quality control
qPCR	quantitative Polymerase Chain Reaction
REVEL	Rare Exome Variant Ensemble Learner
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristic
rRNA	Ribosomal Ribonucleic Acid

SBE	Smad-Binding Element
SBS	Sequencing-By-Synthesis
scRNA-seq	Single Cell RNA-seq
SIFT	Sorting Intolerant From Tolerant
SIMD	Single-Instruction Multiple Vectorized
SMEM	Super Maximal Extended Match
SMRT	Single-Molecule Real-Time
SNV	Single Nucleotide Variant
snoRNA	Small Nucleolar RNA
SW	Smith-Waterman
TAUH	Tampere University Hospital
TCGA	The Cancer Genome Atlas
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TSS	Transcription Start Site
UV	Ultraviolet Light
VEGF	Vascular Endothelial Growth Factors
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WPCM	Weighted Positional Count matrix
ZMW	Zero Wave Guide Detector

ORIGINAL PUBLICATIONS

Publication I Laitinen VH, **Rantapero T**, Fischer D, Vuorinen EM, Tammela TL; PRACTICAL Consortium, Wahlfors T, Schleutker J. 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. 2015 May 15; 136(10):2316-27. doi: 10.1002/ijc.29276. Epub 2014 Nov.

Publication II Määttä K*, **Rantapero T***, Lindström A, Nykter M, Kankuri-Tammilehto M, Laasanen SL, Schleutker J. Whole-exome sequencing of Finnish hereditary breast cancer families. Eur J Hum Genet. 2016 Jan; 25(1):85-93. doi: 10.1038/ejhg.2016.141. Epub 2016 Oct 26.

Publication III Ampuja M*, **Rantapero T***, Rodriguez-Martinez A*, Palmroth M, Alarmo EL, Nykter M, Kallioniemi A. Integrated RNA-seq and DNase-seq analyses identify phenotype-specific BMP4 signaling in breast cancer. BMC Genomics. 2017 Jan 11; 18(1):68. doi: 10.1186/s12864-016-3428-1

Publication IV **Rantapero T**, Wahlfors T, Kähler A, Hultman C, Lindberg J, Tammela TL, Nykter M, Schleutker J, Wiklund F. Inherited DNA Repair Gene Mutations in Men with Lethal Prostate Cancer. Genes (Basel). 2020 Mar 14;11(3). pii: E314. doi: 10.3390/genes11030314

* Equal contribution

1 INTRODUCTION

The year 1990 marked the beginning of a new era in the field of biomedical research as the human genome project was launched. As the sequencing technology was still in its infancy, the project took 13 years to finish. However, during this time array based high-throughput technologies revolutionised the field. These technologies made it possible to study vast quantities of data comprising almost complete transcriptomes and large number of sites of known genomic variations from multitudes of samples at the same time. Still, these methods could not completely replace traditional sequencing methods due to the limitations of the hybridisation based capturing technology.

During the early 2000´s novel sequencing technologies based on massively parallel sequencing reactions were developed. This technology known as Next-Generation Sequencing (NGS) made it possible for the first time to sequence whole genomes in a single run. The introduction of the novel NGS technology in turn gave rise to novel applications, which allowed not only the study of DNA-sequence itself but also various transcribed RNA products and regulatory elements in the genome. The new possibilities introduced by NGS applications culminated in perhaps one of the most ambitious endeavor in biomedical research since the human genome project, the ENCODE project. This ongoing project has identified vast number of gene regulatory elements and networks uncovering the true complexity of genomes and the process of gene regulation.

The advances in the sequencing technology have strongly influenced cancer research. Large consortiums such as The Cancer Genome Atlas (TCGA) have now sequenced thousands of cancer genomes. The characterisation of the transcriptomic and epigenetic profiles of primary tumours has led to the discovery of new tumour types commonly referred to as molecular subtypes. Moreover, novel germline variants, which predispose to cancer and somatic driver mutations, have been discovered. These developments in cancer research have led to a more profound understanding of not only how cancer develops and progresses but also led to the discovery of novel targets for therapeutic intervention.

To be able to harness the vast amounts of data produced by the new sequencing technologies, there is a need for active development of novel bioinformatics tools. Moreover, these tools need to be assembled into seamless bioinformatics analysis frameworks for efficient and accurate data analysis.

The cost of sequencing has been gradually decreasing, which allows for larger sample sizes. Moreover, this has resulted in more complex study designs in which multiple sequencing applications have been combined to study several aspects of molecular biology simultaneously. Therefore, both the ability to manage and analyse large quantities of data, as well as integrate different different data types has become an active area of research in bioinformatics.

2 REVIEW OF LITERATURE

2.1 Next-generation sequencing and its applications

2.1.1 Introduction to sequencing technology

Sequencing enables the characterization of the nucleotide sequence of DNA. The beginning of the modern DNA sequencing era began in 1977, when Frederick Sanger and his colleagues developed a sequencing method currently known as Sanger sequencing. This technique is based on the use of chain terminating nucleotide analogs which, when combined with fluorescent dyes, enabled the automated detection of bases using computers (Sanger, Nicklen, and Coulson 1977). Following this, the development of whole genome shotgun sequencing made it possible to study whole genomes and finally in 2001 the Human Genome Project (Venter et al. 2001) released the first human genome assembly. The automated Sanger sequencing is still used to conduct small sequencing projects as well as validation of results in larger scale projects. However, due the limited amount of sequence, which can be processed using this technology, studying large genomes is time consuming and expensive.

Since the year 2006, novel sequencing technologies started to emerge which made it possible to sequence large quantities of DNA-fragments in parallel. These technologies sometimes called massive parallel sequencing or high-throughput sequencing are nowadays generally known as Next-generation sequencing. Currently, there exists dozens of NGS sequencing technologies of which the most well-known and already commercialised are the Oxford Nanopore, Pacific Biosciences, Illumina and Thermo Fisher Ion Torrent platforms. Different sequencing platforms have varying characteristics and therefore have different targeted areas of applications (Metzker 2010).

The process of sequencing by NGS involves three general steps: Preparing the source material, library preparation and the sequencing itself. The source material being used and its preparation is highly dependent on the sequencing application. The source material can be either genomic DNA or RNA molecules and the

preparation step may involve an additional step for capturing sequence based on the type or genomic location of interest. During the preparation of the source, material the nucleotide sequences will be cut down to smaller fragments for the library preparation. As an additional step the RNA molecules have to be first converted into complementary DNA (cDNA) before fragmentation. During the library preparation, the nucleotide sequences are further prepared for the sequencing instrument. This step is heavily dependent on the sequencing technology being used. Some technologies require large amounts of source material and therefore the fragments need to be amplified. Finally, during the sequencing step sequencing instruments provide the sequence data as identified segments of DNA, which are generally referred to as reads. Figure 1 illustrates generic Illumina/Solexa sequencing workflows for Whole genome sequencing, RNA-seq and Dnase-seq.

2.1.2 Current standard technologies

2.1.2.1 Illumina/Solexa

The Illumina/Solexa technology utilises solid-phase amplification in the library preparation, which involves ligation of adaptors to the ends of the DNA fragments, which are referred to as templates. The adaptors include priming sequences, which hybridise the templates with primer sequences, that are attached to the slide and initiate the Polymerase Chain Reaction (PCR). During this step, the primer sequences attached to the slide will be extended and become complements of the hybridised templates. The template is then removed from its attached counterpart and washed away leaving only the sequence, which will be amplified using bridge amplification. During bridge amplification the extended primer sequences hybridise with their neighbouring unextended primer sequences which initiates the extension of these sequences. This step is repeated eventually leading to formations of clusters of copies of the original template sequence (Metzker 2010).

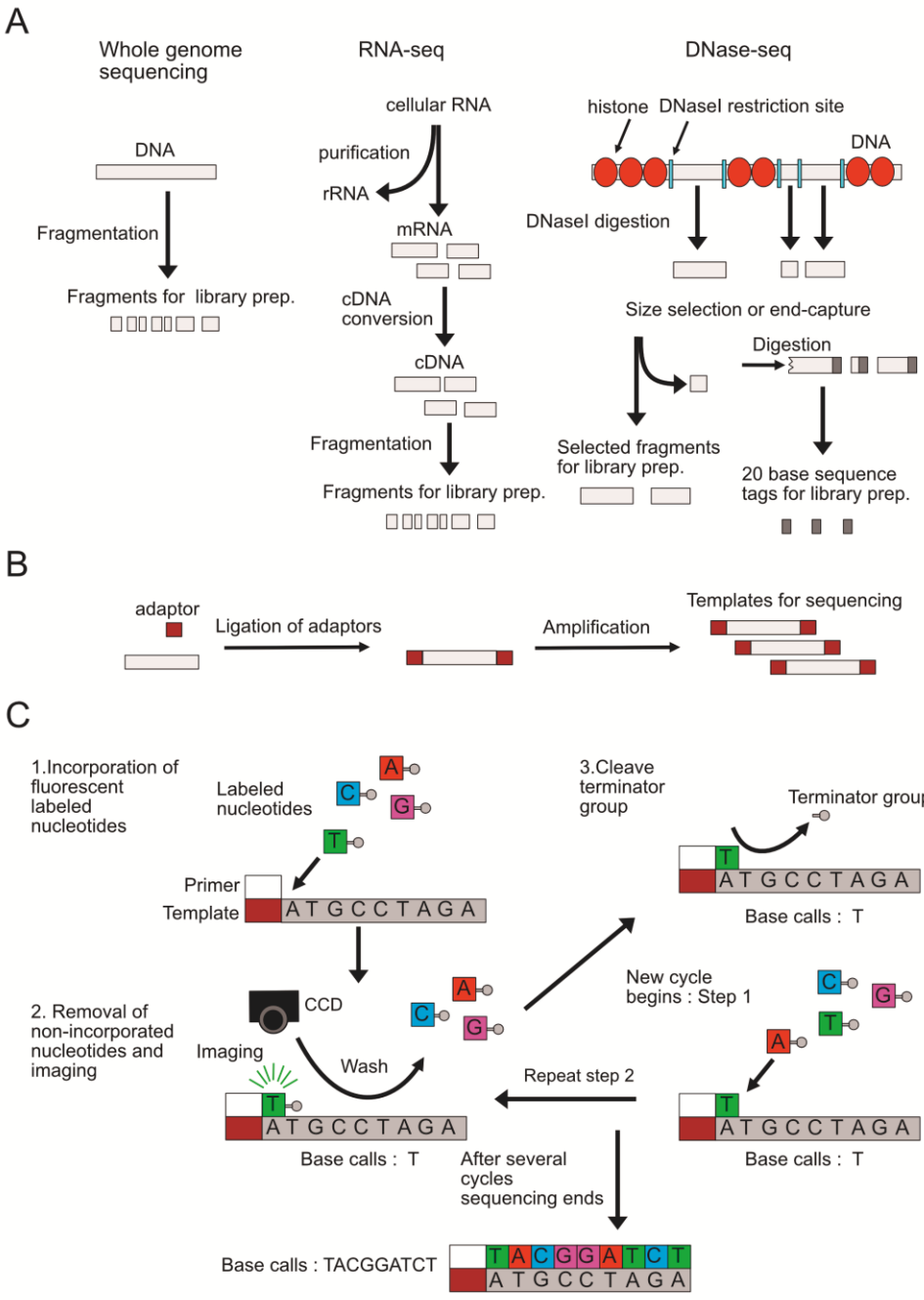

Figure 1. Example of Illumina sequencing workflow for Whole Genome. RNA-seq and DNase-seq. A, Preparation of source material. B, Ligation of adaptors and amplification. C, Sequencing using cyclic reversible termination.

The sequencing procedure is based on sequencing by synthesis (SBS), in which cDNA is synthesised using the DNA-fragments as templates. Typically, the sequencing occurs in a cyclic manner consisting of three steps. During the first step nucleotides are introduced and incorporated to the growing strand followed by the second step in which all the unbound nucleotides are washed away and the identity of the nucleotides are detected. The elongation is paused until the final step, in which a new cycle is initiated (Metzker 2010; Goodwin, McPherson, and McCombie 2016). The Illumina platform applies cyclic reversible termination (CRT) in which fluorescent labeled nucleotides are incorporated by DNA-polymerase one at time to the elongating strand. The nucleotides include termination groups which prevent the addition of more nucleotides during the cycle. After the elongation step the nucleotides which were not incorporated are washed away followed by the imaging using a Couple Charged Device camera (CCD). During the final step the termination group is removed to allow further elongation during the next cycle (Metzker 2010).

The advantages of Illumina sequencing technology over other methods is an extremely low error rate. Illumina has a wide selection of machines which vary in amount of throughput, run times and per base cost. This enables the use of Illumina sequencing platforms in a very broad range of studies from targeted resequencing to large scale whole genome sequencing (WGS) projects (Goodwin, McPherson, and McCombie 2016; Reuter, Spacek, and Snyder 2015).

2.1.2.2 Life Technologies/Thermo Fisher/Ion Torrent

Similar to the Illumina sequencing technology, Ion Torrent is based on SBS. The library preparation involves clonal amplification based on emulsion-PCR. Similarly to the Illumina protocol, the universal adaptors are first ligated to the ends of the fragments which include the priming sequences needed to initiate the PCR reaction. Next, the DNA-fragments are separated from each other and captured into beads by hybridising them to primer sequences on the surface of the beads. The conditions during the capturing step ensure that only one template molecule is captured by a bead. The primers on the surface of the beads are extended based on the template molecule followed by dissociation of the template. This process is repeated using the same template molecule leading to a formation of thousands of copies of the template on the surface of the bead. After the amplification step the beads are then distributed into micro wells for sequencing (Metzker 2010).

The sequencing is based on the detection of the change in pH, which is caused by the release of H+ when a nucleotide is incorporated to the elongated DNA-

strand. The pH change is proportional to the number of incorporated nucleotides. The identification of the nucleotide is possible because only one type of nucleotide is introduced at a time. Once one or more nucleotides have been incorporated, the DNA synthesis halts and new cycle begins by introducing another nucleotide (Reuter, Spacek, and Snyder 2015).

The main advantage of this technology is that it does not rely on optical detection of the bases, which dramatically reduces the cost and run time of the sequencing. Currently, there are two types of sequencing machines available. The first released sequencing machine: Ion PGM has a relatively small throughput and is therefore best suited for targeted resequencing projects or studying small genomes. The most recently released Ion Proton has a significantly higher throughput and can be used for whole exome and transcriptome sequencing. Both sequencing machines produce short reads and therefore are poorly suited for de novo assembly of genomes. The most common error types of this technology are insertion and deletions. Ion Torrent is particularly prone to errors in genomic regions including homopolymers longer than 6 bases, because the pH change does not correlate perfectly with number of incorporated nucleotides (Reuter, Spacek, and Snyder 2015).

2.1.3 Emerging technologies

2.1.3.1 Pacific Biosciences

The Pacific biosciences platform is based on Single-molecule real-time (SMRT) sequencing. After DNA fragmentation, the templates are prepared by adding single-stranded hairpin adapter sequences to the ends of the fragments, resulting in capped templates. The sequencing methodology is based on sequencing by synthesis but unlike the other methods aforementioned, the dye-labelled nucleotides are continuously incorporated to the primer sequence by the DNA-polymerase, which are attached to the bottom surface of Zero wave guide detectors (ZMW). The detectors record the identity of the nucleotides in real-time as they are incorporated to the primer. The platform utilises a strand displacing polymerase, which allows the same template to be sequenced multiple times increasing the accuracy of the base calls. Furthermore, because of this the amplification of template molecules can be avoided (Reuter, Spacek, and Snyder 2015).

The strengths of this technology are short run times and longer read lengths which are on average >14kb but can reach to lengths of 60kb. Since the PCR

amplification step is not required, the method is also less prone to GC bias compared to the methods relying on template amplification. Nevertheless, because single templates are sequenced, the error rates are higher compared to the methods mentioned previously. Since the errors are distributed randomly, the accuracy of the consensus base call can be improved by increasing the coverage or multiple passes around the same template. Still, the high per base sequencing cost and the lower throughput compared to the methods producing short reads hinders its use in large-scale genome studies. Nonetheless, this technology is highly suitable for de novo assembly of small genomes as well as in resequencing projects in which the aim is to improve the current genome. This technology has also been used in the detection of large structural variants as well as in the study of differential isoform usage (Reuter, Spacek, and Snyder 2015).

2.1.3.2 Oxford Nanopore

Nanopore sequencing represents an alternative for previously described methods, which rely on SBS. The general principle of this sequencing method is that DNA fragments or individual nucleotides are transferred through a small channel. The nucleotide passing through the channel is identified based on its induced change in current, which is unique for each type of nucleotide. The current Oxford Nanopore sequencing platform is comprised of hundreds of independent micro wells, which include synthetic bilayers perforated by nanopores. The template preparation consists of DNA-fragmentation and ligating two adapter sequences. Since the method does not rely on fluorescently labeled nucleotides, the template amplification step is optional. The first adapter is bound to a motor protein and a molecular tether whereas the other adapter is a hairpin oligonucleotide, which is coupled with a so-called HP motor protein. The sequencing is driven by the motor proteins, which transfer the DNA-templates through the nanopores (Reuter, Spacek, and Snyder 2015).

Because of the library design both strands of the template molecule can be sequenced which improves the accuracy of the base calls significantly. The strengths of the Oxford Nanopore technology are the minimal library preparation steps and the flow cell design which together allows for small sequencing devices. Moreover, the sequencing machines produce extremely long reads. However, currently the low throughput and high error rates limit the use of this technology for studying larger genomes. Therefore, this technology has mainly been used to study small organisms such as bacteria and yeast. (Reuter, Spacek, and Snyder 2015)

2.2 Applications of NGS technologies

2.2.1 Genome analysis

The most typical application of NGS is the characterisation of the genomes which can be divided into two main categories: "De novo sequencing" and "re-sequencing". De novo sequencing is typically used to sequence an unknown or a small organism in order to assemble its genome whereas re-sequencing is commonly used to sequence an organism with a known reference genome to characterise variation in the genome. Genome sequencing methods can be further categorised based on whether the full genome is being sequenced (Whole genome sequencing) or specific portions of the genome are captured for analysis (targeted sequencing and Whole Exome Sequencing).

2.2.1.1 Targeted sequencing

Targeted sequencing is an application of NGS in which only selected regions of the genome are being sequenced. The target regions are first captured and then fragmented for library preparation. There are several methods for target capturing which can be divided into three main categories: Hybrid, selective circularisation and PCR amplification capture. In the hybrid capture, the target regions are captured by hybridisation with complementary nucleic acid sequences, also known as probes, in a solution or on a solid support. Selective circularisation involves single stranded probe sequences, which contain a stretch of universal sequence flanked by target specific sequences. The target specific sequences are complementary to the sequences flanking the target genomic site and during the capture hybridise with these regions. Subsequently, the gap between the target specific sequences is closed by gap filling reactions and finally ligation of the loose ends results in circular nucleic acid molecules containing the regions of interest. In PCR amplification capture, PCR is used to selectively amplify the target regions by using complementary primer sequences of the flanking regions of the target (Mertes et al. 2011).

Targeted sequencing is more cost effective in comparison to WGS and thus used for studies in which only specific regions are of interest. The capturing methods differ in many respects such as the maximum size of the target region which can be captured, required amount of input DNA, the enrichment of reads obtained from the target and cost efficiency. Although careful selection of the capture method

based on the above mentioned parameters enables usage of targeted sequencing in wide variety of applications, targeted sequencing has its shortcomings. The major issue is the relative unevenness of the coverage of reads which can cause difficulties in downstream bioinformatics analyses such as variant calling (Mertes et al. 2011).

2.2.1.2 Whole-exome sequencing

Whole exome sequencing (WES) is a special case of targeted sequencing in which the target consists of exonic regions. The exonic regions are captured using hybrid capture described previously. The most widely used methods for exome capture are sold as commercial kits, including SureSelect (Agilent) TruSeq Capture (Illumina) and SeqCap EZ (Roche NimbleGen) but also custom methods have been applied and developed (García-García et al. 2016).

WES is faster and more cost effective compared to WGS although the price gap has narrowed drastically during the past years (Hayden 2014). Thus, WES is more scalable which enables better statistical power by sequencing more samples. In addition, the amount of data being produced is much smaller which makes the data more manageable, further reducing the expenses by limiting the computational infrastructure required to analyse the data.

Whole exome sequencing is however limited as it requires a well-annotated organism in order to design the probes for the exonic regions. Other disadvantages of WES include less uniform coverage and a more profound allele distribution bias compared to WGS, resulting in less accurate variant and genotype calls (Lelieveld et al. 2015). Finally, the most obvious disadvantage of WES compared to WGS is that regions outside exonic regions, such as regulatory regions, cannot be studied.

Despite its shortcomings and due to its cost effectiveness WES is a widely used sequencing application. In general, WES is best suited for large-scale population studies of the exotic regions of well-known species. Studies of human Mendelian diseases represent one such example since the variants associated with the disease phenotype are known mostly to occur in the exonic regions (Bamshad et al. 2011).

2.2.2 Transcriptome analysis

Transcriptome analysis involves the characterisation of the sequences being transcribed by an organism including both coding transcripts, such as mRNAs, and non-coding transcripts such as lincRNAs, snoRNAs and miRNAs to name a few.

Perhaps one of the most commonly utilised techniques for studying the transcriptome is RNA-seq. It can be used to study sufficiently long transcripts while short transcripts (< 200 bp) can be studied using other specialised methods such as small RNA-seq.

2.2.2.1 RNA-seq

RNA-seq enables the profiling of the entire transcriptome including protein coding genes as well as non-coding transcripts such as lincRNAs and repetitive elements. Compared to the earlier array-based methods no prior knowledge of the transcriptome is required, which allows detection of novel isoforms and non-coding transcripts which are transcribed only in specific tissues or conditions. Furthermore, it is possible to assemble the whole transcriptome for organisms for which a reference genome has not yet been constructed. Other advantages over the previous array-based technologies include a higher dynamic range of detected expression levels as well as more accurate estimation of the abundance of different transcript isoforms (Wang, Gerstein, and Snyder 2009).

In RNA-seq the total RNA content is first extracted from the sample followed by removal of ribosomal RNA (rRNA) using either poly-A capture or rRNA depletion. The purified RNA is then converted to cDNA. Sequencing templates are then prepared by adding adaptor sequences to either one or both ends of the fragments depending on the library preparation strategy. Subsequently, the templates undergo the standard sequencing steps required by the sequencing platform being used (Wang, Gerstein, and Snyder 2009).

RNA-seq is one the most popular sequencing applications as it provides a very comprehensive view of the transcriptome. However, a major drawback of traditional RNA-seq is that it cannot reveal heterogeneity within a sequenced sample, which can contain multiple cell types. Instead, the obtained abundance estimates for the transcripts reflect the average abundances over the populations of the cells. Recent developments in sequencing technologies have now made it possible to study the transcriptomic profile on a single cell level. This technology is referred to as single cell RNA-seq (scRNA-seq) and it is now being widely used to study areas of research, which is beyond the capabilities of traditional bulk RNA-seq (Saliba et al. 2014).

2.2.3 Epigenome analysis

The epigenome is comprised of all chemical changes occurring in DNA and histones, which together make up the chromatin. Numerous sequencing techniques have been developed to study the different aspects of epigenetic modification in the genome. For studying methylation, techniques such as Bisulfite sequencing (BS-seq) and Methylated DNA immunoprecipitation sequencing (MeDIP-Seq) have been developed. Histone modification can be studied using techniques such as ChIP-seq whereas the overall accessibility of the genome can be investigated using DNase-seq and ATAC-seq.

2.2.3.1 DNase-seq

The nucleus DNA is organised into a structure called the chromatin. Its' basic unit is a nucleosome which consists of approximately 146 bp of DNA wrapped around a histone octamer. The organisation of the DNA as nucleosomes primarily serves as a way to condense the chromatin in order for it to fit inside the nucleus. In addition, the density of the packaging, which varies across the genome, also plays a significant role in gene regulation. The densely packed regions, referred to as heterochromatin, are generally inaccessible to transcriptional machinery and thus genes located within these regions are not expressed. Contrary to heterochromatin, genomic regions, which are less densely packed, are generally known as euchromatin. These regions are more accessible to proteins involved in gene regulation and transcription and therefore genes located within these regions are often actively expressed. The chromatin structure is dynamic and regulated by changes in the composition of the histone proteins composed of octamers and by different post-translational modifications of the histone tails (Valencia and Kadoch 2019).

DNase I is an endonuclease, which digests double stranded DNA by preferentially cleaving the phosphodiester bonds adjacent to pyrimidine nucleotides. The genomic regions which are nucleosome depleted are sensitive to digestion because the chromatin is exposed and allows the binding of DNase I(Weintraub and Groudine 1976). These regions are generally referred to as DNase I hypersensitive sites (DHS). In contrast, regions tightly packed around nucleosomes and other higher order structures are highly resilient to digestion (Elgin 1981). Because open chromatin regions are accessible to various regulatory proteins, these regions are likely to harbour active genetic regulatory sites including promoters, enhancers, silencers, insulators and locus control regions. Since these regions often coincide

with DHS sites, methods for capturing these sites have been developed as early as from the late 70s´ (Song and Crawford 2010).

The early methods used to study the DHS sites suffered from low throughput and therefore their application has been limited. Nevertheless, since the development of NGS technologies, the old methods for capturing DHS sites have been coupled with the novel sequencing techniques to generate novel sequencing applications. One of these techniques is known as DNase-seq, which can be used to characterise the DHS sites across the whole genome (Song and Crawford 2010). Moreover, this technique allows the detection of transcription factor binding sites (TFBS) within DHS regions. What makes it possible is the fact that similar to nucleosomes, transcription factors (TF) can protect the chromatin from digestion at the genomic site where they are bound. This can be observed in the data as lowered accessibility within DHS sites also referred to as TF footprints (Kaplan et al. 2009). There exists two widely used protocols for DNase-seq. In both protocols cells are first lyzed to release the nuclei, followed by the digesting of the genomic DNA using the restriction enzyme DNAseI. In the "double hit" protocol small fragments of between 50–100bp are selected for by using gel electrophoresis whereas in the "end-capture" protocol the ends of all DNA-fragments are ligated to specially designed linkers followed by MmeI digestion yielding 20 bp tags. Depending on the protocol, the fragments or tags are then sequenced by the standard NGS protocols (Sabo et al. 2006; Song and Crawford 2010).

DNaseq has been proven to be a valuable tool in the ENCODE project for characterising the regulatory elements in various cell lines (Dunham et al 2012). However, this technique requires large amounts of input DNA, because much of the DNA is lost during the purification steps. This limits its usefulness, especially when studying clinical samples (Sabo et al. 2006; Song and Crawford 2010). Recently, single cell DNase-seq also known as Pico-seq has been developed which can be used to study heterogeneity within samples and requires less input DNA (Jin et al. 2015). Moreover, during the past decade, other sequencing technologies designed for mapping DHS sites have been developed including ATAC-seq, FAIRE-seq, MNase-seq, and NicE-seq. Particularly ATAC-seq has gained popularity because the low amount of required DNA and the possibility to study nucleosome displacement in high resolution (Chang et al. 2018).

2.3 Methods for analysing NGS data

The general analysis workflow of NGS data can be roughly divided into three steps: Sequence data quality control and preprocessing, read mapping or genome assembly, and downstream analysis. During the quality control and preprocessing step the quality of the sequencing data is assessed and reads can be trimmed or filtered out before the alignment step. After quality control and preprocessing, the reads are commonly aligned against a reference genome or alternatively assembled as a genome. Finally, depending on the sequencing application appropriate downstream analysis is performed. Figure 2 shows an illustration of the general analysis workflow for sequencing data analysis and examples of typical downstream analysis.

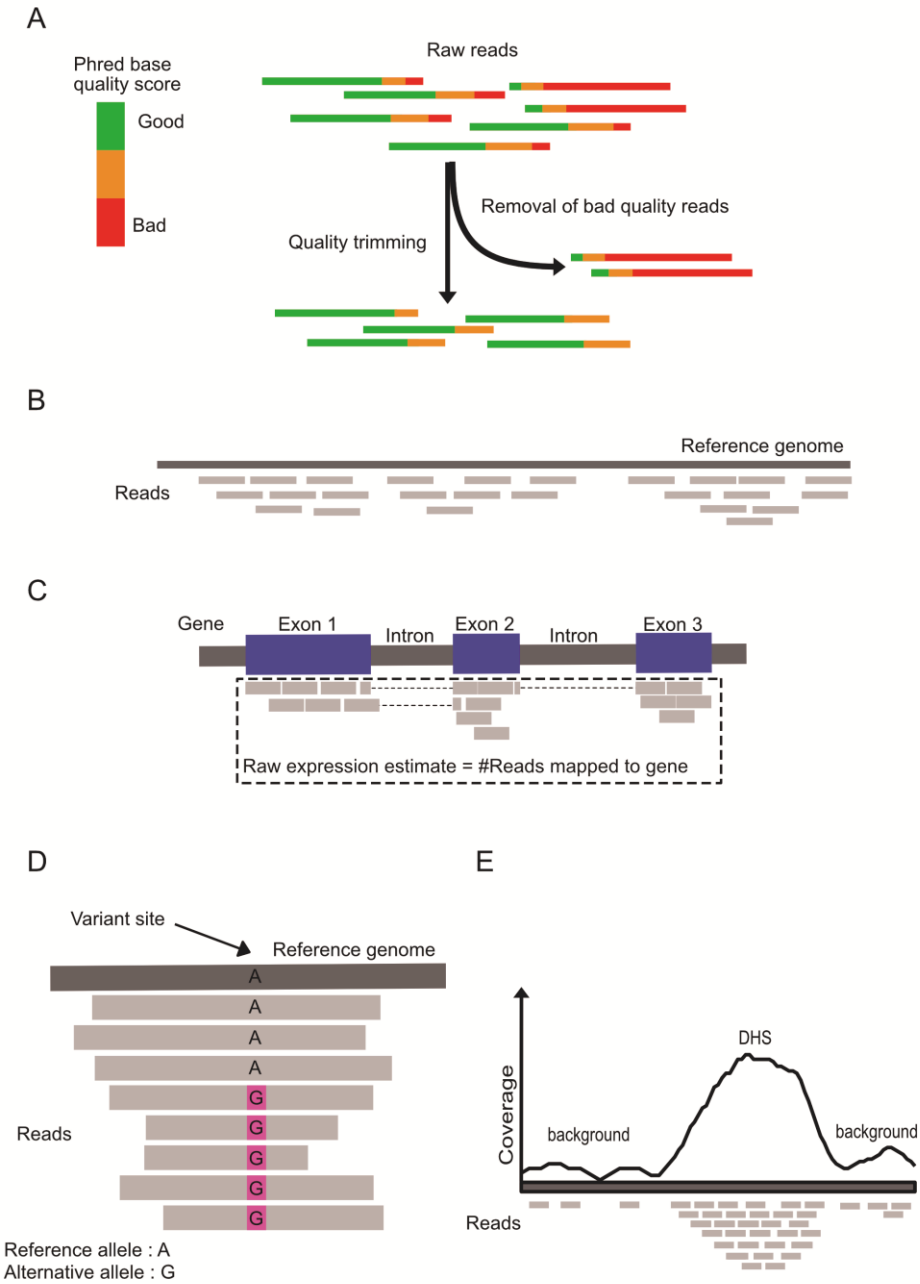

Figure 2. General workflow for the sequencing data-analysis. A, Quality control and pre-processing. B, Alignment against reference genome. C, Gene expression quantification (RNA-seq) D, Variant calling (WGS, WES and Targeted sequencing). E, Peak calling for detecting DHS sites (DNase-seq).

2.3.1 Quality control and data pre-processing

Sequencing data analysis begins with the assessment of data quality to ensure that the experiment has been successful. In addition to monitoring the overall quality, it is also important to detect bad quality samples, which might introduce bias later in the downstream analysis. Furthermore, based on the quality assessment, raw reads can be filtered and preprocessed to improve the accuracy of the results of the downstream analysis (Wang 2016).

The most important quality metric is the phred base quality score (Q), also simply referred to as the phred score, which is a standard metric determined for each base call by the sequencing machine. The phred score is defined according to the following formula:

$$Q = -10log_{10}P$$

,where, **P** is the probability of an erroneous base call**.**

Typically, a phred score of 20 is considered the minimum for the base quality score to be considered reliable which corresponds to a probability of 1 % that the base call is incorrect (Wang 2016).

Phred scores can be used to calculate summary statistics in order to evaluate the overall quality of the sequencing run as well as filter or trim bad quality reads. One such summary statistic is the average per sequence phred quality score, which is the average phred score calculated across all base calls for a given read. Typically in a successful experiment, the mean of the average per sequence phred quality score is close to 30 (Wang 2016). Reads that have low average per sequence phred quality scores can be filtered out to avoid any bias in the downstream analysis (Guo et al. 2014).

Another phred score based summary statistic that is commonly evaluated is the per base phred quality score. This quality metric is determined by calculating the average phred score across all reads for each sequenced base. This metric is especially important when evaluating the quality of reads produced by platforms based on SBS for which it has been observed that the quality of the base calls begin to systematically drop towards the end of the read (Wang 2016). It is a common practice to trim reads such that bases, which have a phred score lower than 20 are removed from the ends of the reads. This procedure will improve the mapping or assembly of the reads (Guo et al. 2014).

Overrepresented sequences are typical artifacts caused by remnants of adaptors or barcodes used for multiplexing. To improve the read mapping or assembly,

adaptors and barcode sequences are typically removed from the reads (Guo et al. 2014). Moreover, in the case of sample contamination it is possible that frequently occurring sequences originate from a completely different organisms. The reads identified originating from contaminants are removed to avoid any bias in the downstream analysis. (Zhou et al. 2013). In addition, the purity of a sample can be evaluated using the per sequence GC content. The GC content of the reads should be normally distributed in the case that reads come from a single organism. Therefore, in the case that this distribution deviates from normal it may indicate a presence of contamination (Guo et al. 2014).

It is also important to monitor the amount of duplicated reads. The main source of duplicated reads is the PCR amplification step which is required by some of the sequencing platforms such Illumina or Ion torrent platforms (Metzker 2010). The quantity of duplicated reads can be considered as measure of the complexity of the sequencing library. In general the more complex the library, the better we are able to characterise the sample (Bansal 2017).

Finally, it is common to evaluate the nucleotide composition of the reads. Ideally, the frequencies of the four nucleotides should be approximately equal at each position of the read. Nevertheless, in some of the sequencing applications such as RNA-seq, systematic biases are typical and therefore should not be considered as sign of a bad quality sample (Guo et al. 2014).

Apart from the previously mentioned general quality metrics, different sequencing platforms have also their own specific metrics. For example, in Illumina sequencing machines the flow cells are organised as tiles. Per tile sequencing score indicates the sequencing quality scores in different tiles of the sequencing machines. Large differences in the base quality score between tiles might indicate issues occurred during the sequencing process caused by air bubbles or debris in the flow cells (Robinson et al 2017).

A wide variety of tools for evaluating different Quality Control (QC) metrics exist of which FastQC is one of the most popular (Andrews 2010). Commonly used methods for the preprocessing steps involving quality and adapter trimming tools include, for example, Cutadapt and Trimmomatic (Bolger, Lohse, and Usadel 2014; Martin 2011). Furthermore, there are a number of dedicated tools for removal of reads originating from contaminant species such as DeconSeq and Fastq screen (Schmieder and Edwards 2011; Wingett and Andrews 2018). Finally, the recently developed tools such as fastp and afterQC integrate the calculation of QC metrics and the aforementioned preprocessing steps such as quality and adaptor trimming, filtering of bad quality reads and removing contaminants. These methods have been

shown to improve the performance in terms of speed in comparison to the earlier single purpose methods (Chen et al. 2017, 2018).

2.3.2 Read alignment

Read alignment or mapping refers to the process of determining the position of the genome were the read originated from. Typically, this step is done after preprocessing and quality control of the reads. Because sequencing produces millions of reads, the read mapping process is a highly computationally intensive task for organisms having large genomes such as human. Moreover, sequencing machines produce erroneous base calls which need to be taken into account when mapping the reads to the genome. Therefore, a huge amount of effort has been spend on developing algorithms which accurately map the reads to the reference genome but at the same time are computationally efficient.

2.3.2.1 The principles of read alignment algorithms

Because of sequencing errors, read mapping can be considered as an approximate string matching problem. To solve this problem, the current mapping software apply two main principles: filtering and indexing. The main idea of the filtering is to limit the search space by excluding regions, where the read could not have originated from. For memory efficient filtering the reference genome or alternatively the reads are stored into specific data structures, which are generally referred to as indices (Reinert et al. 2015).

The most common filtering approaches utilise either so-called pigeon hole lemma or shared q-gram counts. The pigeon hole lemma states that if a read with exactly k errors is divided into k + 1 non-overlapping pieces, also known as seeds, at least one of the seeds will not contain an error. The mapping algorithms that operate based on this principle try to find exact matches in parallel for each of the k + 1 seeds by scanning the reference. All found exact matches are considered as the candidate regions, which will be further investigated during the following seed-extension phase (Reinert et al. 2015).

Q-gram is defined as group of all possible strings of length q over an alphabet which, in the case of read mapping, consists of A, G, T and C. In the q-gram approach the reference is first divided into overlapping regions. Subsequently, for each possible q-gram exact matches are found simultaneously in the reads and the

reference regions. Finally, the candidate regions for each read are selected based on a threshold of number of shared q-grams. This threshold is based on the worst case scenario that k number of errors are equidistantly distributed along the read. According to the q-gram lemma this results in n - (k + 1)q -1 required shared q-grams between the read and the candidate regions, where n is the length of the read (Reinert et al. 2015).

The efficient utilisation of q-grams requires a data structure called the q-gram index, which is implemented using two tables: occurrence table and a lookup table. The occurrence table holds the positions where a specific q-gram occurs in the read. The q-grams are organised such that the positions of q-grams, which occur multiple times in the read, are stored as consecutive entries in the occurrence table. The lookup table is used to retrieve the positions from the occurrence table. This table contains indices related to occurrence table as entries for each q-gram. The query of the position of a q-grams is done first by converting the q-gram to a numerical value c using 4-base system. The lookup table is organised such that this numerical value corresponds to the index holding the information about the q-gram. The entries of the indices c and c + 1 correspond to a half open interval in the occurrence table which contains the positions in which the q-gram occurs in the read. In practice, using a simple lookup table as described above would consume huge amount of memory. Instead read mapping algorithms use more advanced data structures such as hash tables, suffix arrays, enhanced suffix arrays or the Ferragina-Manzini index (FM-index) (Reinert et al. 2015).

After the approximate mapping phase the candidate regions are explored in more depth in a step which is typically referred to as seed-extension. This involves extension of the seed alignments at each candidate region to find the highest scoring local alignment. Several alternative scoring schemes have been introduced in the past. The most commonly used scoring scheme is the Smith-Waterman (SW), which allows user defined penalties for mismatches, gaps and extension of gaps, which makes it very flexible allowing the incorporation of base quality scores, which makes the mapping more tolerant against sequencing errors. Some of the recent methods utilise a more advanced single-instruction multiple vectorised (SIMD) variant of the classical SW algorithm to improve the speed of the local alignment step. After the best local alignments have be determined they are ranked based on their alignment score. The final decision on which of the alignments are accepted is based on user given parameters (Reinert et al. 2015).

2.3.2.2 Read alignment algorithms designed for general purposes

Most of the read mapping algorithms can be used for aligning reads regardless of the sequencing application with RNA-seq being one example of an exception. The read alignment algorithms are also generally applicable to analyse sequencing data from different sequencing platforms but in order to achieve optimal mapping results the parameters of these tools might need to be adjusted to take into account platform specific sequencing errors. Currently, a wide variety of mapping software exists among which BWA and Bowtie2 are some of the most popular.

The original BWA algorithm utilises a backtracking algorithm to find matches for entire reads (end to end) by allowing a user defined maximum number of mismatches and gaps. The accepted alignments are scored based on user defined mismatch and gap penalties and the algorithm reports the alignments with the highest alignment score. As an index, the original BWA algorithm makes use of a suffix array, which has been compressed using the Burrows-Wheeler transform (BWT) (Li and Durbin 2009).

The BWA-SW is a modified version of the original BWA algorithm which makes use of a seed and extension strategy to first find exact matches using a backward search algorithm in a suffix trie representation of the index which is compressed using BWT. The exact matches are then extended using the SW algorithm to find the best alignments. (Canzar and Salzberg 2017). The BWA-MEM is the most recently developed version of BWA, which was developed for short read mapping. The main algorithmic principles of BWA-MEM are the same as those used in BWA-SW. However, instead of using a traditional seed extension method in which matches are found for fixed length seeds it finds so-called super maximal extended matches (SMEMs) which limits the search space more efficiently. The MEMs are exact matches, which cannot be extended in either directions. Reads can have multiple MEMs of which some can be contained by other MEMs. If a MEM is not contained by any other MEM it is considered a SMEM (Ahmed, Bertels, and Al-Ars 2016). In the seed extension phase BWA-MEM utilises a banded SW algorithm to find the best alignments for the determined SMEMs (Canzar and Salzberg 2017).

Similar to BWA-SW and BWA-MEM, Bowtie2 makes use of a seed and extension strategy. Bowtie2 first creates equally spaced seeds from the read and its reverse complement followed by a search of exact matches in the reference. Similar to BWA-SW, Bowtie2 uses the FM-index, which is compressed using BWT. The exact matches are extended using SIMD-accelerated dynamic programming approach,

which allows the introduction of gaps and mismatches to find the best alignments (Canzar and Salzberg 2017).

2.3.2.3 Read alignment algorithms designed for RNA-seq

Most alignment algorithms heavily penalise large gaps in the alignment because large deletions or insertions are relatively uncommon in the genome. However, RNA-seq produces reads originating from spliced transcripts and therefore large gaps need to be allowed in order to properly map reads spanning splice junctions. For this purpose mapping tools generally known as spliced aligners have been developed.

A common strategy applied by many of the early developed spliced aligners such as Tophat and Tophat2, uses two-step process for read mapping. Initially the reads are mapped using conventional read mapping approaches, which are able to map reads that do not span the splice junctions. The clusters of mapped reads define putative exons, which can be used to detect the splice junctions. During the second step the unmapped reads are realigned against the flanking sequences of the putative splice junctions to complete the alignment. (Kim et al. 2013)

One of the more recently developed spliced read mapping algorithms STAR deviates from the two-step strategy by using Maximal mappable prefixes (MMPs). In this approach for each read the maximal mappable section is found starting from the first base of the read. The complete alignment is then found by finding the MMP for the unmapped part of the read. STAR utilises uncompressed suffix array indices which makes the mapping very fast but requires substantial amount of memory (Dobin et al. 2013).

Other more recently developed tools HISAT and HISAT2 utilise two types of FM-indices for the genome: One of the indices holds the whole genome and the others, so-called local indices, are comprised of short segments, which cover the full genome. The consecutive short segments are designed such that they share short overlapping regions to simplify the alignment crossing the boundaries of two local indices. This strategy leads to substantial gains in alignment speed without the cost of requiring additional memory in comparison to the earlier aligners (Kim, Langmead, and Salzberg 2015).

2.3.3 Discovery of germline variants and genotype calling

After aligning the reads to a reference genome each genome position can be screened for the possible occurrence of a variant. Variants can be defined as loci where the sequence deviates from the reference. Each genomic locus can be defined in terms of alleles. The alleles matching the reference are commonly referred to as reference alleles whereas the alleles that do not match the reference are referred to as alternative alleles. Diploid organisms such as humans carry two copies of chromosomes and can therefore carry a variant in both, only in one or none of two chromosomes at the same loci. The concept of a genotype characterises each genomic loci based on the count of reference and alternative alleles. In the case that an individual carries two reference alleles the genotype is considered to be homozygous for the reference allele whereas if the individual carries two alternative alleles the genotype is considered to be homozygous for the alternative allele. In the case that one of the alleles is a reference allele and the other is an alternative allele the genotype is considered to be heterozygous.

In principle the genotype at a specific genomic locus can be inferred from the sequencing data by counting the frequencies of called bases from the reads that have been aligned to that position. In the case of diploid organisms the frequencies of the reference and the alternative allele (assuming that only one type of alternative allele can be found) are approximately 0.5 in the case that the genotype is heterozygous as reads can be assumed to be sampled randomly from either copy of the chromosome. In the case of homozygous genotypes for reference or alternative allele only reference alleles or the alternative alleles should be observed respectively. However, because of errors occurring during sequencing and read mapping, calling variants and genotypes solely based on the frequencies of base counts may lead to large numbers of erroneous genotype calls. Due to this fact, sophisticated alignment data preprocessing and variant calling methods are needed to ensure the quality of the variant calls.

2.3.3.1 Alignment data preprocessing

The most important quality metric of the variant calling methods is the phred quality score of each base call. Since low phred scores indicate a high probability of an incorrect call, less weight should be assigned to base calls having low phred scores when determining the genotype. However, it should be noted that the base quality scores produced by sequencing machines are prone to systematic errors and thus

cannot be fully trusted. These erroneous base quality scores can be corrected by performing a preprocessing procedure known as base quality score recalibration (BQSR). Most of the available methods such as the one adopted by the Genome analysis toolkit (GATK) attempt to correct the systematic errors by applying machine learning algorithms to model the errors empirically and then adjusting the quality scores accordingly. These methods require a set of true variants, which can be obtained from databases such as dbSNP and HapMap (DePristo et al. 2011). As these databases are still limited to only few organisms, dedicated tools such as Lacer have been developed to perform BQSR without any prior knowledge of true variants (Chung and Chen 2017).

Another common issue is the amplification bias associated with the short sequencing technologies. The duplicated reads caused by over enrichment of specific fragments can bias the results, as duplication allows the propagation of single erroneous base calls and thus affects the observed base call frequencies. Some variant calling software require duplicated reads to be removed completely while others require them to be marked in order to avoid producing biased results. These procedures can be accomplished using tools such as Samtools and Picard (Broad Institute 2016; Li et al. 2009).

Some methods such as GATK, VarDict and Strelka2 also incorporate indel based realignment in order to improve the detection of indels in cases when read mapping algorithms fail to align reads correctly. Typically, these regions have multiple consecutive SNVs, which lie at the ends of the reads, which are more favored by the aligners over introducing a gap in the alignment. The indel realignment attempts to correct the issue by realigning the reads at these regions and generating the most parsimonious consensus alignment given all the reads (DePristo et al. 2011; Kim et al. 2018; Lai et al. 2016).

2.3.3.2 Variant and genotype calling

The variant calling tools generally apply either heuristic or probabilistic strategies. The heuristic strategy based methods such as VarScan set thresholds for features such as minimum allele counts, read quality, and read depth at the position of the putative variant site, which variants must exceed in order to be called as true variants (Koboldt et al. 2012). In contrast, the probabilistic methods such as GATK and Strelka2 try to estimate probabilities for each genotype, which can be used to not only call the variants but also assess the reliability of the variant calls. The most common statistical model used by these methods is the Bayes´ model, which requires

that the prior probabilities of genotypes, and the conditional probability of observed reads given the genotype are estimated from the data. In general tools apply equal prior probabilities for each genotype as default but also allow user defined values. (Kim et al. 2018; McKenna et al. 2010)

The calculation of the conditional probabilities varies between the different tools. For instance GATK uses a haplotype calling approach in which it first constructs all possible haplotypes based on the reads and then finds sites which deviate from the reference which are considered as the candidate variant sites. Then likelihoods describing the probability of each read given all the possible haplotypes are calculated. These per read likelihoods are then used to calculate the probabilities for each allele in each candidate variant site. Finally, the probabilities of the alleles given the reads can be then used to calculate the conditional probability of the observed data given a genotype. The genotype maximising this probability is considered to be the true genotype. The genotype calls are done for each sample independently but GATK also includes a module for refining the genotype calls, which makes use of the initial genotype calls. (McKenna et al. 2010)

Some methods such as GATK also include post filtering of the initial variant calls in which additional metrics, such as those utilised by the heuristic methods, calculated from the sequencing data are used to filter out possible artefacts. To this end, GATK uses a multivariate Gaussian classifier trained using high quality datasets of known variants. Alternatively, it is possible to filter variants directly using these metrics similar to the heuristic methods (DePristo et al. 2011; McKenna et al. 2010).

2.3.4 Quantification of gene expression from RNA-seq

One of the most common aims of the analysis of RNA-seq data is to quantify gene expression. RNA-seq is well suited for this purpose because the number of reads originating from a gene is comparable to its mRNA quantity at the time the RNA was extracted. Currently, there exists two main strategies for quantifying the expressions of genes from the raw sequencing data. The first strategy performs the quantification in two steps. The reads are first aligned against the reference genome using splice aware alignment tools, which is followed by calculation of the number of reads mapping to each gene. The alternative strategy utilised by so-called pseudo-aligners, involves aligning reads against transcriptome, which enables the quantification directly by calculating the reads mapped to the transcripts (Stark, Grzelak, and Hadfield 2019).

Since genes commonly have multiple isoforms, counting the reads to yield an estimate for the gene expression is a non-trivial task. There are multiple ways to summarise the transcripts to gene level. The early methods such as HTSeq utilise a simple approach in which isoforms of a gene are collapsed into a single composite transcript consisting of representative exons. The representative exons are generated based on user defined rules which correspond to simple set operations. The standard rule is to find the union of all overlapping exons, which will be considered as the representative exon in the composite transcript. This unambiguous representation of transcripts as a composite transcript makes it possible to simply count the reads mapping to each of the exons of the composite transcript and summing them together to yield the abundance estimate for that gene (Anders, Pyl, and Huber 2015).

The counting procedure described above yields reasonable estimates of gene expression but this approach does not apply for quantifying the expression of individual isoforms. Since isoforms have overlapping exons it is not always possible to determine from which isoform the read originated from. Therefore, recently developed methods attempt to model the probabilities from which isoform the read originates from to yield estimates for the isoforms expressions, which can be further summarised to gene level. RSEM is an example of the most popular methods adopting this methodology. It has been shown that not only RSEM is capable of estimating the isoform abundances it also produces more accurate gene abundance estimates. (Li and Dewey 2011)

All the above-mentioned methods rely on a two step approach of first aligning the reads against the genome followed by counting the reads aligning to genes. More recently developed methods utilise a more straightforward approach by aligning the reads directly against the transcriptome. Instead of creating an index of the whole genome the index is constructed of the transcript sequences of all known transcripts. Tools that use this method such as Kallisto and Salmon produce transcript level abundances, which can be summarised to gene level. Similar to RSEM, both of these methods utilise the Expectation maximization (EM) algorithm to yield abundance estimates, which are probabilistic. The major advantages of using pseudo-aligners is that they are extremely fast compared to the earlier approaches because the index is limited to transcripts instead of the whole genome. (Stark, Grzelak, and Hadfield 2019)

2.3.5 Peak detection

Many sequencing applications such as Dnase-seq and ChIP-seq aim to capture selectively genomic regions of interest. Therefore, most of the reads can be considered to originate from the regions of interest. However, in practise the sequence capture is not perfect and errors may also occur during sequencing and read mapping leading to varying amounts of reads randomly distributed among the genome. To identify the regions of interest from the background noise sites having significantly higher coverage relative to the background need to be detected. The procedure is commonly referred to as peak detection, which can be accomplished with the use of dedicated tools.

The general framework for peak detection involves dividing the genome into bins followed by counting the number of mapped reads falling to each bin. Most tools operate on the bin counts directly apart from F-seq, which transforms the bin counts into continuous signal using kernel density estimation. Subsequently, a statistical model is fit to the data in order to yield the background distribution of mapped reads, which can be utilised to detect true DHS sites from the background signal.

Perhaps the simplest statistical framework is utilised by MACS, which assumes that the background distribution follows a Poisson distribution (Zhang et al. 2008). DFilter applies a linear filter known as the Hotelling observer which aims to maximize the ROC-AUC by estimating the mean and covariance of the counts profiles of signal positive and signal negative (background) regions. Both of the above mentioned methods classify each bin into either an enriched or background category. If the count deviates significantly from the background distribution the bin is classified as enriched and otherwise the bin is classified as background. After bin classification, the consecutive enriched bins are merged into peaks, which are interpreted as DHS (Kumar et al. 2013; Zhang et al. 2008).

Because F-seq evaluates continuous signal based on the kernel density estimation it classifies points within a given window instead of bins. In order to classify each point a distribution of kernel estimates corresponding to the background is generated by first calculating the average number of points falling within each window given the data. Then a number points equal to the average number of points within a window are randomly placed within a window followed by calculating the kernel density estimate for a fixed point within the window. As this procedure is repeated multiple times the distribution of density estimates approximates to Gaussian distribution. Given the background distribution each point can be evaluated by comparing the observed true density estimate against a threshold which

is set to S standard deviations above the mean of the background distribution where S is a value set by the user (Boyle et al. 2008).

While most tools utilise only the read counts some methods such as Zinba take into account other covariates such as the mappability of reads and the GC content of the bins. Moreover, Zinba fits a zero inflated negative binomial model, which is used to classify genomic windows to either enriched, background, or regions, which are lacking coverage due to biases present in the sequencing technology as well as read mapping (Rashid et al. 2011).

2.4 The biology of cancer

2.4.1 Hallmarks of cancer

Cancer is the second leading causes of death worldwide. It is a large family of diseases characterised by a series of genetic and epigenetic alterations in a population of cells, which lead to abnormal cell growth and possible invasion, and dissemination of cancerous cells to distant parts of the body. Cancers have been traditionally distinguished from other malignancies by six hallmarks: Sustained proliferative signaling, evasion of growth suppression, invasion and metastasis, replicative immortality, induction of angiogenesis and resistance to cell death. Furthermore, more recent findings have highlighted two additional hallmarks including evasion of immune destruction and deregulation of cellular energy metabolism. It should be noted that these hallmarks are not by any means independent characteristics of cancer but are interconnected by common enzymes and regulatory proteins (Hanahan and Weinberg 2011).

In normal tissue, the number of cells are tightly regulated by limiting the production and release of growth promoting signals. However, in cancer this homeostasis is disturbed. Cancer cells can sustain their growth promoting signalling by producing these signals either by themselves or by stimulating surrounding stromal cells to produce these signals. Moreover, cancer cells commonly increase the abundance of growth factor receptors making them more sensitive to the growth signals. Cancer cells can also become completely independent of external growth signals by adopting structural changes in the growth factor receptors or in proteins occurring in the downstream signalling cascade (Hanahan and Weinberg 2011).

In addition to cell growth promoting factors, the cell proliferation is controlled by growth suppressing factors known as tumour suppressors. Tumour suppressors can inhibit growth by preventing the cells to proceed through the cell cycle and directing the cells to become senescent or undergo programmed cellular death also known as apoptosis. In cancer, loss of function mutations are frequently observed in tumour suppressor genes which allow the malignant cells to pass cell cycle checkpoints as well as avoid programs that would lead the cells to become senescent or undergo apoptosis (Hanahan and Weinberg 2011).

Programmed cell death is an important mechanism for controlling the size of population of cells as well as monitoring the genomic integrity of this population. The decision whether the cell should undergo apoptosis or not is controlled by the balance of the intracellular quantities of BCL2-family members of pro- and anti-apoptotic regulatory proteins. Elevated levels of oncogenic signalling as well as DNA-damage typically triggers apoptosis. However, tumours can develop mechanisms that prevent apoptosis. The first mechanism involves loss of function of key tumour suppressors. Furthermore, tumours can perturb the balance of apoptotic regulator proteins by increasing the expressions of anti-apoptotic proteins or down regulating pro-apoptotic regulators. The third mechanism involves evading apoptosis by increasing the abundance of survival signals (Hanahan and Weinberg 2011).

Normal cells can only replicate a limited number of times. Contrary to normal cells, cancer cells have unlimited replicative potential. Two main protective mechanisms prevent the cells from replicating endlessly. Normally cells will eventually after several replication cycles reach a point where they are forced into a state of cellular senescence, which means that while remaining viable they are no longer able to replicate. Sometimes cells are able to evade the control machinery of the cell, which will lead to a state of crisis. The underlying mechanism behind the crisis is the shortening of telomeres located in the ends of the chromosomes past the point that replication is no longer possible. As a result, the cells will undergo apoptosis. In cancer the telomerase enzyme, whose expression is repressed in normal cells, is commonly active. This enzyme catalyses the extension of telomeres preventing the crisis and thus enabling replicative immortality (Hanahan and Weinberg 2011).

During the progression of cancer, the malignant cells adopt the ability to invade the surrounding tissues and eventually may end up in the bloodstream through which they can reach distant parts of the body where they form distant colonies known as metastases. The complex series of changes leading to these events are commonly

referred to as epithelial to mesenchymal transition (EMT). This phenomenon involves loss of adherens junctions with the neighbouring epithelial cells as well as contact to the extracellular matrix (ECM). This leads to perturbed signaling and eventually changes in cell morphology to resemble typical spindly/fibroblastic characteristics. During EMT, the malignant cells adopt the ability to secrete enzymes capable of degrading the extracellular matrix, which facilitates the invasion the malignant cells to neighbouring tissue. Eventually, cancer cells can move to distant sites by entering into the bloodstream either directly from the tissue of origin or via lymphatic vessels (Hanahan and Weinberg 2011).

Similar to normal tissues tumours are dependent on nutrients and oxygen as well as being capable of disposing metabolic waste and carbon dioxide. This means that tumours need blood vessels to sustain their growth. After the embryonic development of novel blood vessels are only transiently activated because of physiological processes such as wound healing and the female reproductive cycle. Tumours are able to induce the sprouting of new blood vessels also known as angiogenesis by secreting proangiogenic signaling molecules such as vascular endothelial growth factors and fibroblast growth factors. Moreover, the angiogenesis can be also induced by the inflammatory immune cells which have been infiltrated to the tumour mass (Hanahan and Weinberg 2011).

In normal cells under aerobic conditions, energy is produced through glycolysis. The glycolysis produces pyruvate, which is consumed by oxygen dependent citric acid cycle in the mitochondria. In contrast, under anaerobic conditions, pyruvate is converted to lactic acid in the cytosol. In cancer, the energy metabolism is commonly altered such that even under aerobic conditions the glycolysis is the main energy source. This phenomenon is known as the Warburg effect, which has been also observed in rapidly dividing embryonic tissues. To sustain sufficient intake of glucose cancer cells have been shown to have increased number of glucose transporters, which mediate the transportation of glucose from the extracellular space to the cytosol (Hanahan and Weinberg 2011).

The immune system has been shown to be crucial for suppressing the growth of tumours. However, some tumours develop the ability to suppress the immune response and thus evade destruction. While these mechanisms are still widely unknown, two possible strategies have been uncovered. The tumours can directly prevent the action of cytotoxic T lymphocytes and Natural killer cells by secreting immunosuppressive factors such as TGF-beta. Alternatively, under certain circumstances such as occurrence of cell death by necrosis, tumours may attract

inflammatory cells such as regulatory T cells and myeloid-derived suppressor cells, which can suppress the immune response (Hanahan and Weinberg 2011).

2.4.2 The genetic and epigenetic background of cancer development

Three mechanisms that are predisposing to cancer have been identified including environmental factors such as carcinogenic chemicals and UV-light, certain viruses such as the papilloma virus and genetic predisposition. The heritability of cancer varies between cancer types. It has been shown that prostate cancer and breast cancer are among the most heritable cancer types. For prostate cancer (PrCa) the most recent estimate for heritability is 58 % whereas for breast cancer (BC) it is 31 % (Hjelmborg et al. 2014; Mucci et al. 2016)

Cancer is a complex disease in which multiple variants distributed among various number of chromosomal loci have been found to contribute to genetic susceptibility to cancer. Moreover, these variants have been found to be highly specific to different ethnic groups. Since cancer is a common disease it was initially assumed that predisposition is mainly due to common variants occurring in the population. Indeed large Genome Wide Association Studies (GWAS) have identified large numbers of common variants associated with both cancer types which supports the "Common disease, common variant" hypothesis (Demichelis and Stanford 2015). However, these common variants have been shown to have only low to moderate effects on the risk of cancer and therefore do not explain the high incidence of cancer observed in some families. (Benafif et al. 2018; Lilyquist et al. 2018).

It has been long known that family history can be used to predict the incidence of cancers such as PrCa and BC. The estimated risk of being affected for men with a family history of PrCa in a first degree-relative is approximately 2-3 folds higher in comparison to other men (Demichelis and Stanford 2015). In BC, the corresponding increase in the risk has been estimated to be approximately two-fold (Beral V 2001). To search for variants associated with the increased risk in families with a history of cancer, studies utilising linkage analysis and most recently NGS have been conducted which have successfully uncovered new low-frequent and rare variants cancer predisposing variants.

To date, GWAS studies have discovered over 180 loci associated to PrCa. Surprisingly, most of them are located in intergenic regions, which suggests that the variants mediate their effects through gene regulation. Furthermore, linkage analysis and sequencing studies of familial cancer patients have been able to identify low

frequency (1 - 5 % in population) and rare (< 1 % in population) variants which have been shown to contribute to prostate cancer susceptibility. Perhaps the most groundbreaking discovery has been the association of linkage signal found in 17q21-22 to a rare variant in *HOXB13* (G85E) which is currently the only confirmed high risk variant associated with prostate cancer observed widely among different populations (Demichelis and Stanford 2015; Schumacher FR et al 2018; Takata R et al 2019).

Similarly to PrCa, in BC several low risk common variants have been found. To date, over 200 variants have been identified. (Lilyquist et al. 2018; Rivandi, Martens, and Hollestelle 2018; Zhang et al. 2020). Moreover, familial studies have identified high-risk variants in *BRCA1*, *BRCA2*, *TP53*, *STK11*, *CDH1* and *PTEN*. In addition, moderate risk variants have been discovered in *CHEK2*, *ATM*, *PALB2* and *NBS1*. The variants in *BRCA1* and *BRCA2* are clearly the most common of the high-risk variants and thus these genes are now routinely used in genetic screening for evaluating the risk of familial BC (Rivandi, Martens, and Hollestelle 2018). Because of the prominent role of *BRCA* variants in predisposition to breast cancer, the most recent familial studies have mainly focused on characterising the variants contributing cancer risk for patients without known *BRCA1/2* variants.

Recent findings have shown that not only the risk but also the aggressiveness of the disease is modulated by germline variants. Notably variants affecting DNA-repair genes such as *BRCA2* and *ATM* have been associated with the development of more aggressive disease and thus can be used as markers for prognosis (Carter et al 2019; Na et al 2017; Pritchard et al 2016). Moreover, studies have found the similar to somatic mutations certain germline variants can have therapeutic implications. So far deleterious germline variants in *BRCA2* have shown to increase the efficacy of both platinum chemotherapy as well treatment with PARP-inhibitors (Warner et al. 2019).

Even though germline variants can increase the risk of cancer, they rarely can lead to the development of cancer alone. Cancer is ultimately the result of both germline variants already present during embryonic development and randomly occurring somatic mutations, which have been accumulated during lifetime. These somatic mutations can be either gain of function mutations, which allow constitutive activity of growth promoting factors or loss of function mutations of tumour suppressing genes. Moreover, genes can be amplified leading to increased levels of the gene product and thus increased activity or deletions, which lead to loss of gene product and activity (Vogelstein and Kinzler 2004).

Large TCGA pan-cancer studies have shown that cancers of specific type can be classified into subtypes based on their mutational and transcriptional profiles. These

subtypes have been associated with many clinically relevant characteristics such as response to therapies and survival (Berger et al. 2018; Hoadley et al. 2014). Still, the genetic background which is characterised by germline and somatic mutations, is not the only defining factor of the characteristics of cancer. Changes in the methylation profile is a well-known mechanism driving cancer and has been shown to give rise to specific cancer subtypes (Witte, Plass, and Gerhauser 2014). Furthermore, a recent pan-cancer study of the chromatin accessibility landscape of TCGA cohort suggests that methylation is not the only epigenetic factor associated with cancer development. The differentially active regulatory elements defined by the chromatin landscape have been shown to be unique to different cancer types and also define subtypes within the cancer types. Moreover, the unique combinations of active regulatory elements do not affect only the transcriptional profile, but also clinically relevant characteristics such as survival and immunological response (Corces et al. 2018).

2.5 Next-generation sequencing in cancer research

2.5.1 Discovery of coding germline variants associated with cancer susceptibility and aggressiveness

The sequencing methods typically used for discovery of germline variants are WES and WGS. So far whole exome sequencing has been the most popular method due to its cost efficiency. While being significantly more expensive compared to WES, WGS provides a more comprehensive view on the germline variant profile of individuals and will likely replace targeted sequencing and WES in the future as it becomes more affordable.

Regardless of the sequencing method being used, the data analysis involves an analysis workflow from raw sequencing data to variant calls. Currently, the GATK best practices protocol is considered the golden standard for human germline variant discovery. Because of the variant discovery step thousands of variants are typically obtained (Van der Auwera et al. 2013; DePristo et al. 2011).

During the following downstream analysis steps, variants which are unlikely to be causal for the disease are filtered out. Contrary to GWAS, which aims to find common low penetrant variants, the resequencing studies focus on discovery of rare variants, which are highly penetrant (Bomba, Walter, and Soranzo 2017). Therefore,

a typical analysis workflow includes a filtering step in which elatively common variants are omitted for further analysis. In order to estimate the allele frequencies in a population, databases such as gnomAD, ExAC and dbSNP can be used (Karczewski et al. 2017, 2019; Sherry 2001).

The rare variants are further assessed based on the impact at the protein level. Variants that introduce or remove a stop codon, referred to as stop gain and loss variants respectively, are considered high impact variants as they can alter the length of the final protein product leading to loss of function (LOF) or complete degradation of the protein product by nonsense mediated decay (NMD). Moreover, frameshift indels and splice site altering variants are considered high impact since they have the potential to cause extensive alterations in the protein sequence. In general, Missense and disruptive in-frame indels variants are considered the second highest priority as they are able to alter protein sequence only locally (Jalali and Gamieldien 2017).

The effects of missense variants and in-frame indels are hard to predict solely based on the identity of substituted amino acids. For instance, missense variants might have a drastic effect if the amino acid being changed has some special role in the function of the protein. In order to be able to predict whether variants have an effect in silico prediction tools have been developed. This software are commonly known as pathogenicity predictors (Eilbeck, Quinlan, and Yandell 2017). Most of the earlier methods such as SIFT use evolutionary information in the assessment of variant pathogenicity based on the principle that the more evolutionary conserved amino acid, the more likely a change in this amino acid will lead to a dysfunctional protein product (Ng and Henikoff 2001). Other more recent methods such as Polyphen2 take in to account additional sequence based features and utilise machine-learning methods which have been trained using known pathogenic and neutral variant datasets (Adzhubei, Jordan, and Sunyaev 2013). Further development of the pathogenicity predictors has led to the development of so-called ensemble predictors such as REVEL and VEST3, which make use of results obtained from several pathogenicity predictors as features which can be used to make the prediction (Ghosh, Oak, and Plon 2017).

Other type of variants including synonymous variants, which are not associated with splice sites, are generally considered low priority since they typically have low impact on the phenotype. Although there is some evidence that these variants are associated with predisposition to cancer, they have typically low penetrance.

Typically, after these prioritisation steps there are still many candidate variants remaining. In a study conducted by MacArthur et al. 2012 including 185 individuals

having no diagnosed hereditary diseases, approximately 100 high impact variants likely rendering 20 genes to be dysfunctional were found. These findings illustrate the challenge in detection of causal variants related to diseases, as certain genes can be dysfunctional and still not lead to the development of a disease. Even if the variant is associated with a disease, it might not be associated with the condition being studied. One strategy to limit number of variants is to focus on genes that has been previously associated with cancer utilising databases such as OMIM, ClinVar and COSMIC. Moreover, ClinVar can be also used to discover variants that are known to be pathogenic and associated with cancer (Landrum et al. 2018; Amberger 2015; Sondka et al. 2018).

2.5.2 Studying gene dysregulation in cancer

2.5.2.1 Finding association of variants and gene regulation (eQTL-analysis)

Previous GWAS-studies investigating the predisposition to cancer have highlighted variants located in non-coding regions of the genome. These findings have led to the conclusion that these variants might affect cancer predisposition by altering the regulation of genes (Nica and Dermitzakis 2013). This hypothesis is supported by a recent study conducted by Corces et al. 2018 in which association between the chromatin structures at the loci of previously reported cancer associated GWAS variants and transcriptional profile of tumours was found.

Variants which have the potential to alter gene expression are called expression Quantitative Trait Loci (eQTL) and are classified into two categories based on their mechanism of action. Variants, which alter gene expression of local genes, are generally referred to as cis-eQTLs, whereas variants, which regulate distant genes, are called trans-eQTLs. Cis-eQTLs are thought to mainly act by altering the binding efficiency of regulatory proteins in promoter and enhancer regions associated with a nearby gene. Currently, the mechanism of trans-eQTLs is less well understood. However, recent studies suggest that they might act indirectly by altering the regulation of nearby genes similar to cis-eQTLs, which would then act as regulators of the more distant target gene (Nica and Dermitzakis 2013).

The relationship of genomic variants and gene regulation can be studied using an approach called eQTL-analysis. Assuming that a variant is biallelic leading to three possible genotypes the eQTL-analysis finds variants associated with the expression of a gene by testing if the expression of the gene differs between any of the groups

defined by the three genotypes. To determine the genotype status and gene expression estimates for the genes of interest, current methodology involves combining whole genome or targeted sequencing and RNA-seq (Majewski and Pastinen 2011).

To test the hypothesis that a variant is associated with the expression of a gene, several computational methods have been suggested. The most commonly used method is to apply simple linear regression, which models the quantitative trait e.g. the gene expression to be dependent on a single variable namely the genotype of a variant site. The linear model is defined by the following formula:

$$g = \alpha + \beta s + \epsilon$$

,where **g** denotes the expression of the gene and **s** denotes the genotype. The estimated parameters **α** and **β** are the intercept and slope respectively and **ε** is the error term, which is assumed to be follow a standard normal distribution.

In this model the genotype groups are numerically coded as 0, 1 and 2 such that 0 group represents the homozygous genotype in respect to the major allele, 1 represents the heterozygous genotype and 2 represents the homozygous genotype in respect to the minor allele. The hypothesis of this test setting is the following:

$$H_0 : \beta = 0, \text{Genotype is not associated to expression}$$
$$H_1 : \beta \neq 0, \text{Genotype is associated to expression}$$

The hypothesis is tested by calculating either t-test, F-test or likelihood ratio test-statistic depending on the number of genotype groups being evaluated. This model assumes that the effect of the genotype is additive such that the gene expression is increasing or decreasing linearly as more minor alleles are introduced at variant site (Shabalin 2012).

Apart from the simple linear model, methods such as ANOVA and mixed models, which can incorporate non-additive effects, have been applied (Lee 2018). Furthermore, non-parametric tests such as the extensions of Mann-Whitney U-test including Kruskal-Wallis have been suggested. These test are generally more robust to outliers in comparison to the parametric test which assume normal distribution and therefore are more suited in experimental settings in which the sample size is small (Qi et al. 2014).

Typically, when conducting an eQTL analysis vast amount of variant-gene pairs are being tested which leads to a requirement of multiple test correction by adjusting the p-values. However, standard methods used for correction of p-values such as bonferroni yield overly conservative results considering the fact that nearby variants

tend to have highly correlated genotypes due to linkage disequilibrium (LD). Therefore, dedicated methods for p-value adjustment have been proposed. Commonly used approach is to determine empirical FDR values such that the original test statistic is compared against distribution of test statistics, which have been calculated for repeated random permutations of the dataset (e.g permutation testing). Because permutation testing is computationally intensive, the more recently developed methods attempt to approximate the FDR estimates by utilising dimension reduction techniques or by modeling the LD structure (Davis et al. 2016; Johnson et al. 2010).

One of the major challenges in eQTL studies is that usually the samples cannot be collected from the tissue of interest but rather the sample collection is limited to peripheral blood, which is easy to obtain. Since the chromatin landscape and methylation status among other factors affecting the transcription profile vary across different tissue and cell types the results obtained from one tissue cannot be directly generalised to others (Nica and Dermitzakis 2013).

To address this challenge publicly available datasets of chromatin structure, histone modification and known transcription factor binding sites as well as gene expression data in various cell/tissue types have been utilised to assess if the findings also apply in the tissue under study. Moreover, the computational methods such as Position Weight Matrix (PWM) matching have been applied to evaluate if the variant has a potential to alter the binding efficiency of regulatory proteins such as transcription factors (Huo et al. 2019; Zhang et al. 2018).

2.5.2.2 Studying the association of chromatin structure landscape and gene regulation

Recent studies have integrated techniques for chromatin state characterisation including ATAC-seq and DNase-seq to RNA-seq to explore the relationship of chromatin structure and gene regulation (Miyamoto et al. 2018; Sieber et al. 2019). The standard workflow for DNase-seq or ATAC-seq involves the detection of open chromatin regions, which are annotated based on their genomic context to uncover putative active enhancers and promoters as well as other regulatory elements. In study designs involving multiple conditions or other distinct groups of samples, typically differentially open chromatin sites are determined in order to characterise the different epigenetic profiles between the groups. Finally, the data is integrated with the gene expression data obtained from RNA-seq in order to find associations between the observed epigenetic and transcriptional profiles.

The integration of chromatin accessibility and gene expression data is highly non-trivial. In the case in which multiple groups have been studied, the differentially open chromatin sites can be associated with their target genes by limiting the candidates to the differentially expressed genes. However, the amount differentially open chromatin sites typically exceeds the number of differentially expressed genes by a large extent. This is due to the fact that multiple open chromatin sites can be involved in the regulation of a single gene. This makes it difficult to associate the regulatory elements to their target genes because the sample sizes are generally too small to find strong correlation between regulatory elements and gene expression. One strategy to address this issue is to make use of the fact that regulatory elements are commonly associated with their nearest gene. An overrepresentation analysis of gene ontology terms or biological pathways can be then done for these genes using standard tools for the purpose such as DAVID to gain insight on the relationship of the chromatin and gene expression profiles (Huang, Sherman, and Lempicki 2009). The shortcoming of assigning the active regulatory elements to their nearest gene is that some of the true associations between regulatory elements and target genes might be missed. The reason for this is that regulatory elements can come to contact from distant region depending on the three dimensional configuration of the chromatin. In order to avoid this issue tools such as GREAT, which take into account groups of genes within a user specified window, can be used to refine the enrichment analysis leading to better understanding of the relationship to between the chromatin dynamics and gene regulation (McLean et al. 2010).

3 AIMS OF THE STUDY

The aim of this thesis was to develop a computational framework for the analysis of Next-generation sequencing data to gain more understanding on genetic and epigenetic background of cancer. Moreover, the focus of this thesis was to study how integration of data originating from different NGS applications can be leveraged in studying the complex regulatory mechanisms in cancer. Specific aims for the study are the following:

1) Develop a framework for annotation and prioritisation of variants to identify germline variants that are associated with cancer susceptibility and aggressiveness. (Studies 1, 2 and 4)

2) Extend the framework to integrate variant and epigenetic data with gene expression data and apply the extended framework to accomplish two subaims:

    1) Discovery of non-coding germline variants, which have a modulatory effect in cancer through, altered gene regulation. (Study 1)

    2) Characterisation of the role of the epigenetic profile in differential BMP4 response in breast cancer and the discovery of transcriptional regulators involved in the process. (Study 3)

4 MATERIALS AND METHODS

4.1 Study subjects and materials (1, 2, 4)

In studies, 1 and 2 the ancestry of all the study subjects was Finnish while in study 4 the ancestries were either Finnish or Swedish. The cohorts used in studies 1, 2 and 4 are summarised in Table 1.

4.1.1 Familial prostate cancer patients (1, 4)

The prostate cancer samples have been collected by the Laboratory of Cancer Genetics in the University of Tampere and Tampere University Hospital (TAUH). Moreover, Finnish Cancer Registry and church parish registers have been utilised for identification of additional patients and their first-degree relatives. In studies 1 and 4 only the most representative families have been selected for the analysis such that each family had either at least three affected family members or two affected family members which were either first degree relatives or at least one of them had been diagnosed with prostate cancer before the age of 60 years.

In study 1, 37 members of the aforementioned families showing a linkage to chromosomal regions 2q37, 17q11.2-q22 or both were sequenced. For the validation with genotyping and association analysis, only the index patients of families were used. In addition, more patients and unaffected relatives were analysed for the co-occurrence of identified genetic variants with the disease phenotype. In study 4, two distinct groups of familial prostate patients were studied. The first group defined as "lethal cases" consists of patients who died of prostate cancer before the age of 65, whereas the second group of samples, defined as "unselected cases", consisted of familial prostate cancer cases, which were not selected based on lethality or aggressiveness.

4.1.2 Sporadic prostate cancer patients (1, 4)

The population based collection of patients with unknown family history of the disease has been collected by the Department of Urology in Tampere University Hospital (TAUH). This collection has been restricted to Pirkanmaa region. The associated clinical data for these patients have been obtained from hospital records. In study 4, the sporadic patients were further selected and classified into "lethal" and "unselected" groups using the same criteria as with the familial samples described above.

4.1.3 Unaffected population control individuals (1)

The blood samples of anonymous voluntary healthy male donors between ages 18 to 65 were obtained from the Finnish Red Cross Blood Transfusion Service. The individuals were all healthy at the time the samples were drawn. In addition, various subsets of unaffected male and female family members belonging to the prostate cancer families included in the study 1 were included to investigate the co-occurrence of the identified variants with the disease phenotype.

4.1.4 High risk HBOC patients from Tampere region (2)

The study subjects were recruited from the Tampere University Hospital Genetics Outpatient Clinic (Tampere, Finland). The individuals with breast and/or ovarian cancer were reviewed based on hospital records and pedigree information. A total of 120 individuals who had strong family history of breast and/or ovarian cancer, which fulfilled the high-risk hereditary BC criteria, and had tested negative for BRCA1/BRCA2 mutations previously identified in the Finnish population, were selected from the recruited group of individuals. The high-risk hereditary criteria were defined as follows. 1) The individual or her first-degree relative was diagnosed with breast or ovarian cancer before reaching 30 years; 2) or two first degree relatives in the family were diagnosed with breast and/or ovarian cancer at younger than 40 years of age; 3) or three first-degree relatives had been diagnosed at younger than 50 years of age; 4) or at least four first-degree relatives had been diagnosed with breast and/or ovarian cancer at any age; 5) or the same individual had breast and ovarian cancer; or 6) male BC was observed in the family. 84 out of the 120 families gave a written consent for participating in the further studies. The individuals were

considered index cases and the recruitment was then further extended to healthy and affected relatives of these families. The cancer diagnoses for the index and the other recruited individuals were confirmed from the hospital records and/or Finnish cancer registry and the pedigrees structures were confirmed based on data collected from the Population Registry center.

4.1.5 High risk HBOC patients from Turku region (2)

The individuals were recruited from the Turku University Hospital Department of Clinical Genetics. The subjects were selected based on the previously described criteria for hereditary breast cancer risk. Furthermore, these patients had been tested to be BRCA1/2 mutation negative according to protocol designed by Turku University Hospital Department of Clinical Genetics. Similarly to recruitment of patients from Tampere region, those individuals whose families gave a written consent for participation to further studies were selected as index patients and recruitment was then extended to their healthy and affected relatives. Hospital records were used to confirm the cancer diagnoses of the index patients and their recruited relatives.

4.1.6 Breast cancer patients with and without ovarian cancer (2)

Breast cancer patients and patients with breast and ovarian cancer were BRCA1/2-negative females of Finnish origin. Formalin-fixed paraffin-embedded (FFPE) breast tissue block samples of breast cancer and breast and ovarian cancer patients were obtained from Auria Biobank (Turku, Finland).

4.1.7 Male breast cancer patients (2)

Forty-four male BC samples were part of cohort which has been described in previously published studies (Syrjäkoski et al. 2003, 2004). In addition, five patients were recruited from the Turku University Hospital Department of Clinical Genetics (Turku, Finland), as described in the previous section.

4.1.8 Unaffected population control individuals (2)

Unaffected control individuals were female or male donors whose blood samples were obtained from the Finnish Red Cross. The donors´ blood samples have been collected from the Tampere, Turku and Kuopio regions. The donors were healthy, anonymous volunteers whose age ranged from 18 to 65.

4.1.9 Swedish lethal prostate cancer patients (4)

The Swedish patients were collected as part of Cancer of Prostate in Sweden (CAPS) study, a population-based case-control study, from four of the six regional cancer registries which cover the entire population of Sweden (for more details see Lindmark et al. 2004). The candidate subjects for the study were selected based on pathologically or cytologically verified adenocarcinoma of the prostate. The physician treating these subjects were contacted and asked for approval for the patients to participate to the study. If permission was given the physicians were asked to mail a letter describing the study and asked the patient to send a reply letter to the administrator at the cancer registry. Those subject who gave a written consent for participation were selected as part of the cohort. The clinical data was retrieved from the Cancer Registry or from the National Prostate Cancer Registry. The same definition used to determine the lethal Finnish patient cohort was applied to the Swedish prostate cancer cases.

4.1.10 Ethical aspects (1, 2, 4)

Written informed consent was obtained from each participant in the studies. In studies 1 and 4 the permission for the Finnish familial PrCa sample collection and the use of data stored in the Finnish Cancer Registry was granted by the Ministry of Social Affairs and Health. Permission to collect and use samples from unselected patients treated at the Hatanpää City Hospital was granted Institutional Review Board of the City of Tampere. In study 2, permission to collect data from high-risk HBOC families and to use the data from the Finnish Cancer Registry and Population Register Centre was granted by the National Institute for Health and Welfare. Permission to collect and use blood samples and clinical data from high-risk HBOC who visited the Tampere University Hospital Genetics Outpatient Clinic (Tampere, Finland) was received from the Ethical Committee of Tampere University Hospital.

Furthermore, permission to use blood and clinical tissue samples of deceased individuals for medical research purposes was obtained from the National Authority for Medicolegal Affairs. Permission to collect and use blood samples and clinical data from the high-risk HBOC families who visited the Department of Clinical Genetics, Turku, University Hospital (Turku, Finland) was granted by Ethical Committee of Turku University Hospital. Moreover, the Auria Biobank (Turku, Finland) provided permissions to use their samples. Cancer of Prostate in Sweden (CAPS) study was approved by the ethics committees at the two participating academic institutions, Karolinska Institute and Umeå University. For more details see (Laitinen 2016; Määttä 2016 and Lindmark et al. 2004)

Table 1. Summary of samples included in studies 1, 2, 4.

Sample type	Study	Individuals
Familial prostate cancer patients	1	63/188/243/84[a]
Unaffected male family members	1	3/112/15[b]
Female family members	1	2/92[c]
Sporadic prostate cancer patients	1	1105
Male population controls for prostate cancer	1	923
Tampere HBOC individuals	2	14[d]/65[e]
Turku HBOC individuals	2	10[d]/64[e]
Healthy relatives belonging to HBOC families	2	13
Male breast cancer patients	2	49
Female population controls	2	989
Male controls	2	909
Finnish lethal prostate cancer patients	4	47
Swedish lethal prostate cancer patients	4	75
Finnish unselected PRCA patients	4	70
Finnish population controls (ExAC)	4	3,307
Swedish population controls	4	6,192

a Targeted re-sequencing/Sequenom validation/Co-segregation analysis/RNA-seq
b Targeted re-sequencing/Co-segregation analysis/RNA-seq
c Targeted re-sequencing/Co-segregation analysis
d Sequenced using WES
e Genotyped using Sanger sequencing or TaqMan SNP genotyping assays

4.1.11 Cell lines (3)

Breast cancer cell lines BT-474, HCC-1954, MCF-7, MDA-MB-231, MDA-MB-361, MDA-MB-436, and T-47D as well as the normal immortalised mammary gland cell line MCF-10A were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA). For RNA-seq and DNase-seq, one sample per cell line and

treatment was used. For qRT-PCR, samples representing three biological replicates were collected at indicated time points and pooled.

4.2 Methods

4.2.1 Data preparation

4.2.1.1 Cell culture and treatments (3)

The cells were cultured according to the manufacturer's recommendations. The authentication of the cell lines was done by genotyping and were regularly tested for mycoplasma infection. Cells were seeded and allowed to adhere for 24 h after which they were treated with 100 ng/ml recombinant human BMP4 protein (R&D Systems, Minneapolis, MN, USA) or vehicle (BMP4 dilution solution). Samples were collected 3 h after the treatment.

4.2.1.2 Targeted DNA re-sequencing (1)

The targeted re-sequencing of the prostate cancer associated loci 2q37 and 17q11.2-q22 was done at the Finnish Institute for Molecular Medicine Finland (FIMM), University of Helsinki. SeqCap EZ Choice array probes (Roche NimbleGen, Madison, WI) were used for capturing the target regions and the sequencing was done using Genome Analyzer IIx (Illumina, San Diego, CA) platform.

4.2.1.3 Whole exome sequencing (2, 4)

In study 2, the sequencing was done by BGI Genomics Institute (Hong Kong) using SureSelect Human All Exon 51M kit (Agilent technologies) and HiSeq 2000 platform. The sequencing steps followed the protocols of Agilent, Illumina and BGI. The quality control and data preprocessing was done according to standard protocols of BGI. In study 4, the lethal prostate cancer samples were sequenced by SciLifeLab using HiSeq 2500 instrument. The targeted exome capture was done using Agilent SureSelect Human All Exon (V5) following the standard protocol. The 45 unselected

PRCA cases were sequenced as part of multinational ICPCG study at Mayo Clinic, Rochester, MN, USA. Exome capture was performed using Agilent 36Mb (V1) 50Mb (V2) and V4+UTR SureSelect Human. Samples were pooled post-capture and sequenced three to a lane using Illumina HiSeq 2500 instrument. The 25 HPC family were sequenced by BGI Genomics Institute (Hong Kong) using HiSeq 2000 instrument. In the exome capture SureSelect Human All Exon 51 M kit (Agilent Technologies, Inc., Santa Clara, CA, USA) was used following the protocols by Agilent, Illumina and BGI.

4.2.1.4 RNA-seq (1, 3) and DNase-seq (3)

In studies 1 and 3, RNA sequencing was performed at BGI Genomics institute (BGI Hong Kong) using Illumina HiSeq2000 sequencing platform following the standard BGI protocols. Similarly, the DNase-seq library construction and sequencing steps were carried out at the BGI Genomics institute according to their standard practice. Sequencing was performed using the Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA, USA).

4.2.2 Data analysis

The tools and databases utilised in the data-analysis are listed in Table 2.

4.2.2.1 Quality control, read alignment, variant calling and annotation of targeted sequencing data (1)

The quality control, data preprocessing, read alignment and variant calling were performed according to FIMM's Variant-Calling Pipeline (Sulonen et al. 2011). Only variants, which were shared by all affected family members, were selected for further analysis. These variants were annotated against Ensembl version 65 gene set, which was retrieved from the UCSC Genome Browser (Fujita et al. 2011; Flicek et al. 2013) using in house python scripts. The variants effect on the phenotype were assessed with three in silico pathogenicity prediction programs: MutationTaster, PolyPhen2 and PON-P (Adzhubei I, Jordan DM, Sunyaev SR 2013; Olatubosun et al. 2012; Schwarz et al. 2010). In order to assess the effects regulatory potential of the non-coding variants, these variants were queried against Regulome database

(RegulomeDB) (Boyle et al. 2012). Population allele frequency data for the remaining variants was retrieved from dbSNP (Sherry 2001). Common variants with minor allele frequencies above 0.05 were filtered out. In order to prioritise the variants based on their host genes, a list of known PrCa-associated genes was collected from the COSMIC and DDPC databases (Forbes et al. 2011; Maqungo et al. 2011). Furthermore, pathway data was gathered from Pathway Commons, KEGG and WikiPathways. In addition, Gene Ontology (GO) data was retrieved from Ensembl BioMart version 65 (Cerami et al. 2011; Gene Ontology Consortium 2004; Ogata et al 1999; Pico et al. 2008; Smedley et al. 2009). Only those variants, which were located in genes previously, linked to PrCa or genes which were functionally similar to PrCa-associated genes were chosen for the validation.

4.2.2.2 Validation of variants with genotyping and testing for association (1)

Validation was performed on germline DNA from 2,216 subjects, including 1,293 cases and 923 population controls. Of the affected individuals 1,105 represented unselected PrCa patients from the Pirkanmaa Hospital District, Tampere, Finland and the remaining represented 188 index cases from Finnish HPC families 10 were included in the study. The control DNA samples represents anonymous male blood donors, which were provided by the Finnish Red Cross Blood Transfusion Service. Genotyping was performed at the Technology Centre. The validation was done in 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) and the genotype calls were also checked manually to ensure the quality of the calls. Genotyping quality was examined using a QC procedure, which involved success rate checks, duplicated samples and water controls. PLINK and R-software were used to perform Hardy-Weinberg equilibrium (HWE) tests and calculating the odds-ratios. (Purcell et al. 2007; R Core Team 2013). Two sided Fisher's exact test was used to test for association of variants to PrCa.

4.2.2.3 eQTL mapping and data analysis (1)

The eQTL analysis was based on the RNA-seq data and on the SNV genotypes obtained from targeted DNA sequencing. In order to focus discovery of variant gene

associations, which are likely to be associated with PrCa, two alternative approaches, were used to pre-filter either the target genes or variants prior to eQTL analysis. In the first approach, only genes that were located in 2q37 and 17q11.2-q22 and were differentially expressed between PrCa cases and controls were included in the eQTL analysis. For these genes, variants located within 1MB up- or downstream were considered as potential cis-regulatory variants and selected for eQTL analysis. In the second approach further referred to as "modified cis-eQTL", large genotype dataset from the iCOGS study (Eeles et al. 2013) including 2,824 unselected Finnish PrCa patients and 2,440 controls was utilised to identify PrCa associated variants. To test for association Fisher´s exact test was used. Variants having a p-value less than 0.005 were considered significant. Finally, those variants located in 2q37 and 17q11.2-q22 showing association to PrCa, which were also found in the cohort analysed by targeted sequencing, were selected for eQTL analysis. The eQTL analysis was performed separately for these two pre-filtered sets of variants utilising generalised Mann-Whitney test implemented in the R-package GenomicTools (Fischer et al 2017). The significance level for variant-gene associations was set to p-value = 0.005.

RegulomeDB was used to annotate and assess the regulatory potential of the detected eQTLs. The ENCODE datasets (Dunham et al 2012) were retrieved from the UCSC Genome Browser website for visualisation purposes using the table browser tool (Karolchik et al. 2004). As a general indicator of regulatory potential, we used the dataset that contained enriched DNase hypersensitive sites in 125 cell types. Moreover, to elucidate the regulatory potential of eQTLs in PrCa, we used the LNCaP DNase datasets containing DNase hypersensitive sites in LNCaP cells under normal and androgen-induced conditions (Thurman et al. 2012). TF-binding site data were obtained from the Txn Fac ChIP V3 dataset, which contains ChIP-seq experimental data on 91 cell types and 189 TFs.

4.2.2.4 Quality control, read alignment and variant calling of whole exome sequencing data (2, 4)

In study 2, the reads were aligned with Bowtie2 (Langmead and Salzberg 2012) against the hg19 reference genome. PCR duplicates were removed using Samtools (Li et al 2009) and reads aligned with mapping quality less than 10 were filtered out. Variant calling was done using bioinformatics toolkit Pypette with default parameters (Annala 2016). In study 4, the reads were aligned using BWA against the hg19 reference genome (Li and Durbin 2009). The PCR-duplicates were marked using PICARD (Broad institute 2016) and the base score recalibration was done using

GATK (McKenna et al. 2010). Variant calling was performed using GATK following the GATK best practises protocol for germ-line exome-sequencing data (DePristo et al. 2011; Van der Auwera et al. 2013). Likely false positive variants were filtered using the variant quality score recalibration procedure implemented in GATK by setting the tranche threshold 99.0. Furthermore, variants having an allele fraction of less than 0.3 or a coverage of less than 12 were filtered out. Finally, variants with a readPosRankSum less than or equal to -1.7 were discarded. In order to assess sample quality Bedtools (Quinlan 2014) was used to calculate the genome wide coverage for each sample. Those samples that had less than 30% of bases covered by at least 20 reads were excluded from further analysis.

4.2.2.5 Variant annotation and prioritization for validation (2, 4)

In study 2 variants were annotated with Annovar using refseq genes as reference gene set. The pathogenicity of the variants were evaluated utilising pre-computed pathogenicity predictions from several software available for Annovar. Population minor allele frequencies were retrieved from 1000 Genomes (Auton et al 2015), ESP6500 (Fu et al. 2013), SISU (Lim et al. 2014), GME (Scott et al. 2016), Kaviar (Glusman et al. 2011), ExAC (Karczewski et al. 2017) and GnomAD (Karczewski et al. 2019) databases included in Annovar (Wang et al. 2010). The variants, which were likely neutral, were omitted for further analysis. Variants were considered neutral if their minor allele frequency exceeded 0.05 in any population frequency database or variant was synonymous SNV or non-frameshift indel, which did not alter a splicing site. Furthermore, only those variants, which targeted DNA-repair genes, were selected for further validation.

In study 4 the variant annotation was carried out using Annovar. Variants found in DNA repair genes were selected for further analysis. The intergenic and common (MAF > 0.01) variants were filtered out and the remaining variants were classified into two categories: potentially damaging and neutral. The potentially damaging variants were further classified into two categories (Tier 1 and Tier 2) based on their impact. The classification was based on database of reported associations of variants to clinical phenotypes (ClinVar) (Landrum et al. 2018) and two pathogenicity predictors, CADD (Rentzsch et al. 2019) and REVEL (Ioannidis et al. 2016). Moreover, in order to assess the pathogenicity of protein truncating variants (PTVs) the coordinates of known protein domains from the UniProt database were utilised (Bateman 2019). Those variants that were reported as likely benign or benign in ClinVar were classified as neutral. Protein truncating variants (stopgain, frameshift

indels or splicing site altering variants) were classified as Tier 1 variants if they had a CADD phred score equal of higher than 20. Furthermore, the variants were required to be reported to be pathogenic or likely pathogenic by the ClinVar database or alternatively known to affect a protein domain reported in Uniprot. Variant was considered to to affect a protein domain if they were located in position, which occurs before or within a protein domain. All non-synonymous single nucleotide variants (missense variants) reported to be pathogenic or likely pathogenic by ClinVar or had a CADD phred score ≥ 20 and REVEL score ≥ 0.75 were classified as Tier 2 variants.

4.2.2.6 Discovery of genes associated with aggressive Finnish and Swedish PrCa cases (4)

Each DNA-repair gene which harboured at least one tier 1 or tier 2 variant were tested for association to aggressive PrCa by comparing the frequencies of potentially damaging DNA repair gene variant carriers among the lethal PrCa patients to the frequency in unselected PrCa patients and the two control populations (Finnish and Swedish ExAC control). Comparison was performed using two-sided Fisher's exact test, considering P-value less than 0.05 as indication of statistically significant difference between the compared cohorts. Tier 1 and Tier 2 variants were assessed separately.

4.2.2.7 Data analysis of RNA-seq (1, 3)

The quality control was done using FastQC (Andrews 2010). The reads were aligned using Tophat2 (Kim et al. 2013) against hg19 reference genome. The read counts for the genes were determined using HTSeq (study 1) (Anders et al. 2015) and Pypette (study 3) (Annala et al 2016). The raw read counts were normalised by library size using the median of ratios normalisation implemented in DESeq (study 1)(Anders and Huber 2010) and DESeq2 (study 3)(Love et al. 2014) packages for R. In study 1, the differential gene expression analysis was performed using a two-sided Mann–Whitney test with a p-value cutoff of 0.05. In study 3, the genes were considered differentially expressed if the absolute log2 ratio value was 0.75 or greater and the absolute difference in read counts in the two conditions was at least 50.

4.2.2.8 DNase-seq quality control, read alignment and detection of DNase hypersensitive sites

The quality control was performed using FastQC and the reads were mapped to hg19 reference genome using Bowtie2. DNase hypersensitive sites (DHSs) were detected using DFilter (Kumar et al. 2013). The parameters were set according to the developer's recommendation: The standard deviation was set to 2, bin size to 100 bp and kernel size to 50. Moreover, the refine parameter was set to true. In order to mitigate the effects of mappability and coverage bias DNase I input controls were used. All found DHS sites which were covered less than 20 reads in either DNase I treated samples or input controls were considered to be likely false positives and thus omitted from further analysis. Furthermore, DHS sites which were located in positions that overlapped blacklisted regions collected by the ENCODE consortium were filtered out. The adjacent DHSs, which were located within distance 100 bp or less, were merged together. Finally, the merged DHSs were annotated using Bedtools against Gencode Genes version 19 (Harrow et al. 2012).

4.2.2.9 Discovery of differential DHSs

The difference in chromatin hypersensitivity at DHS sites between the two conditions was assessed using DHS change scores (ΔDHS). The DHS change score for i: th DHS was calculated using the following formula introduced by He et al. 2012 :

$$\Delta DHS = \sqrt{\frac{n_i^{treated}}{\sum_{k=1}^{m}(n_k^{treated})/m}} - \sqrt{\frac{n_i^{vehicle}}{\sum_{k=1}^{m}(n_k^{vehicle})/m}}$$

,where **m** is the total number of DHS sites, $\mathbf{n_i^{treated}}$ is the number of reads mapped to DHS site in the treated sample and $\mathbf{n_i^{vehicle}}$ is the number of reads mapped to the DHS site in the vehicle sample.

The genomic sites having ΔDHS equal or greater than 0.20 were considered differential DHSs and were selected for enrichment analysis. Enrichment analysis was conducted using GREAT (McLean et al. 2010) with default parameters. The results were ranked and selected based on the binomial test such that all FDR adjusted p-values were required to be less than 0.05. All categories including less than

10 genes or more than 1000 genes were omitted from the final results in order to filter out very small or overly generic ontology terms.

4.2.2.10 Correlating DNase coverage of TSS and gene expression

All possible transcription start sites (TSS), of protein coding transcripts obtained from GENCODE were extended by 1000 bases to both directions. The coverage was calculated for each of these extended TSS regions. To summarise the coverage on gene level, the weighted sum of the coverages of the TSSs over all the transcripts associated with that gene was calculated. The weights were determined based on the ratio of the estimated expression of the transcript and the total expression of the gene which were determined using RSEM (Li and Dewey 2011). In the case when the gene was not expressed in one of the conditions, the same ratio, which was observed in the other condition, was used. If the gene was not expressed in either condition, the maximum TSS coverage over all the transcript's TSSs was used as the representative coverage of the TSS of the gene. Based on the coverage, the chromatin status of each gene's TSS was classified into two categories: closed or open. A TSS was considered to be closed if its coverage belonged to the 1. Quintile of the TSS coverages of all genes, in that particular cell line and condition. Otherwise, the TSS was considered to be open. Each TSS was associated with the corresponding normalised expression value of the gene, which were obtained by dividing the expression value obtained after median of ratios normalisation by the gene's total exon length.

4.2.2.11 Prediction of transcription factor binding sites

In order to find potential transcriptional regulators of BMP4 response, DHSs overlapping proximal promoters (2000 bp upstream regions) of upregulated genes were scanned with Position Weight Matrices (PWMs). The PWMs were generated from the curated collection of Weighted Position Count Matrices (WPCMs) obtained from HOCOMOCO database (version 9) (Kulakovskiy et al. 2013).

The PWMs were calculated from weighted matrices of positional counts (WPCM) using the following formula previously introduced by Makeev et al. 2003:

$$S_{b,i} = ln\frac{x_{b,i} + aq_b}{(W + a)q_b}$$

,where $\mathbf{x_{b,i}}$ is the positional count of base **b** in the i:th column of WPCM, **W** is the total weight of the WPCM, **a** is the pseudo count defined as **ln(W)** and $\mathbf{q_b}$ is the background frequency of base **b** calculated across all the analyzed sequences.

The score for transcription factor binding match (**Mj**) was obtained for each position within the peaks by scanning the sequence using the previously defined PWMs. The score for position **j** when scanning with PWM **S** of length **w** is calculated as follows:

$$M_j = \sum_{i=0}^{w-1} S_{b_{(i+j)},i}$$

The PWM was considered to be a match if the PWM score had a p-value less or equal than 0.001. The score thresholds corresponding to the used p-value cut-off were determined using MACRO-APE (Vorontsov et al. 2013).

4.2.2.12 Finding enriched and depleted transcription TFBS in promoters of upregulated genes in the BMP4 stimulated cells

The promoters of upregulated genes were tested for enrichment for transcription factor binding sites by calculating the ratio of enrichment by dividing the observed number of binding sites found for a specific transcription factor across the DHSs of the promoters by the expected number of binding sites for that transcription factor. The number of expected binding sites was estimated based on a background model, which was generated by selecting the DHS sites of all proximal promoters, which were not included in the set of promoters of upregulated genes. To calculate the expected number TFBS the background set was first scanned for TFBS followed by dividing the number of TFBS by the cumulative length of the DHS sites being scanned. Finally, the rate of observed TFBS in the background set was multiplied by the cumulative length of the DHS sites of the upregulated promoters to yield the expected number of TFBS.

4.2.2.13 Co-localization enrichment analysis of selected TFs and known consensus SMAD4-motifs

Selected TFs were tested for co-localization with six known Smad-binding elements (SBEs) including: CAGACA, GTCT, CAGC, CGCC, GGCGCC and GCCGnCGC. The TF and the Smad binding element were considered to be co-localized if the TFBS occured within 200 bp distance of the consensus motif. The observed co-localized TFBSs were compared against expected number of co-localization events calculated for a background set consisting of 200 bp promoter sequences including a match of the consensus motif. The p-values were obtained using the binomial test.

Table 2. Tools and databases used in studies 1-4.

Tool/database	Application
1000 Genomes	Database of genomic variants collected from various sequencing projects and populations
Annovar	Tool for annotating variants
Bedtools	Toolkit for performing operations for genomic intervals
Bowtie2	Tool for short read alignment
BWA	Tool for short read alignment
CADD	In silico pathogenicity predictor for indels and missense variants
ClinVar	Database of known associations of variants to clinical conditions
COSMIC	Database of cancer driver genes and somatic mutations found in various cancer types
dbSNP	Database of genomic variants
DDPC	Database of genes associated to PrCa
DESeq (R-package)	Tool for performing differential expression analysis
DFilter	Tool for detection of DHS sites
Ensembl Biomart	Tool for retrieving gene relatead data from Ensembl
ESP6500	Database of genomic variants collected from various sequencing projects
ExAC	Database of genomic variants collected from various sequencing projects and populations
FastQC	Tool for running sequencing data quality control
GATK	Toolkit for processing alignment data and variant calling and filtering
Gencode genes	Database of genes and transcripts identified by the ENCODE project

Table 2. Continued

Tool/database	Application
Gene Ontology	Database associating genes to biological processes, their molecular function and cellular components
GenomicTools (R-package)	R-package designed for non-parametric eQTL analysis and multidimensional scaling
GME	Great middle eastern database of genomic variants
gnomAD	Database of genomic variants collected from various sequencing projects and populations. Successor of ExAC
GREAT	Tool for enrichment analysis of genomic regions
HOCOMOCO	Database of Weighted Position Count Matrices for transcription factors
HTSeq	Tool for estimation abundance of transcripts on gene level
Kaviar	Database of genomic variants collected from various sequencing projects and populations
KEGG	Database of biological pathways
MACRO-APE	Tool for estimation of p-values for a PWM scores
Mutation taster	In silico pathogenicity predictor for missense variants
Pathway Commons	Database of biological pathways
PICARD	Toolkit for processing alignment data
PLINK	Toolkit for population genetics
PolyPhen2	In silico pathogenicity predictor for missense variants
PONP	In silico pathogenicity predictor for missense variants
Pypette	Toolkit for variant calling
R	Tool for statistical computing
RegulomeDB	Database and web application for priotitising non-coding variants based on regulatory potential
REVEL	In silico pathogenicity predictor for missense variants
RSEM	Tool for estimation abundance of transcripts in gene and isoform level
Samtools	Toolkit for processing alignment data
SISU	Finnish database of genomic variants
UCSC genomebrowser	Integrative database including data from various data sources such as ENCODE, RefSeq and Ensembl
UniProt	Database including protein related data
Wikipathways	Database of biological pathways

5 SUMMARY OF THE RESULTS

5.1 Fine-mapping of 2q37 and 17q11.2-q22 loci in HPC families (1)

5.1.1 Novel variants associated with PRCA predisposition at 2q37 and 17q11.2-q22 loci

A total of 68 individuals belonging to 21 HPC families were fine-mapped using targeted re-sequencing. The targets consisted of two chromosomal loci 2q37 and 17q11.2-q22, which have been previously linked to familial prostate cancer (Cropp et al. 2011). The total number of unique variants discovered across all samples by the FIMM variant calling pipeline was 107,479. Initial variant filtering resulted in discovery of 152 predicted pathogenic variants of which 41 were located in 2q37 and 111 in 17q11.2-q22. After the final prioritisation steps, 44 variants were selected for validation with genotyping. In addition, 14 variants were selected among the predicted neutral variants, which were located in genes which have been previously associated with PrCa.

All together 58 variants were validated by genotyping in total of 1,293 affected individuals consisting of 1,105 sporadic cases and 188 familial cases. In addition, 923 unaffected controls were genotyped. Two case-control analysis were conducted in which the cohorts of affected individuals were compared separately against the unaffected controls. The association analysis found total of 13 variants in seven distinct genes to be statistically significantly associated with PrCa (Table 3). Three of the variants were located in *ZNF652*, whereas *HDAC4, HOXB3*, *ACACA* and *MYEOV2* each harboured two variants. The remaining two variants were located in *HOXB13* and *EFCAB13*. Three of the 13 variants were found in the coding regions while the remaining 10 were non-coding variants.

Four of the variants which were significantly associated with PrCa, were observed in both familial and the sporadic cohorts. Two of these variants (rs116890317 and rs79670217) were located in *ZNF652* and the other two were found in *HOXB3* (rs10554930), and *MYEOV2* (rs13411615). The two *ZNF652* variants had the

strongest association with an increased PrCa risk. Among the familial cases, rs116890317 had the most significant association (OR = 7.8, 95% CI 3.0 – 20.3, p-value = 3.3 × 10−5) and also conferred the highest risk of 3.3 (95% CI 1.4 – 7.5, p-value = 0.003) among the sporadic cases. Rs79670217 had the most significant association with PrCa in the sporadic cohort (OR = 1.6, 95% CI 1.2-2.2, p-value = 0.002) and was the second most significant variant in the familial PrCa patients (OR = 1.9, 95% CI 1.2 – 3.1, p-value = 0.009)

Table 3. Statistically significantly associated variants to PrCa in loci 2q37 and 17q11.2-q22.

				Familial cases vs. control		Unselected cases vs.control	
Chr	Variant	dbSNP ID	Gene	P-value	OR (95% CI)	P-value	OR (95% CI)
17	c.-258-3097A>T	rs116890317	*ZNF652*	$3.3*10e^{5}$	7.8 (3.0 – 20.3)	0.003	3.3 (1.4 – 7.5)
17	c.-258-19749A>C	rs79670217	*ZNF652*	0.009	1.9 (1.2 – 3.1)	0.002	1.6 (1.2 – 2.2)
17	c.-105-850_-105-848delTGT	rs10554930	*HOXB3*	0.01	1.4 (1.1 – 1.8)	0.034	1.2 (1.0 – 1.4)
17	c.-371-137_-371-136insA)	rs35384813	*HOXB3*	0.013	1.4 (1.1 – 1.8)	0.073	1.1 (1.0–1.3)
2	c.958G>A, p.Val320Ile	rs73000144	*HDAC4*	0.018	14.6 (1.5 – 140.2)	0.078	5.9 (0.7–47.9)
2	g.241075991A>C	rs13411615	*MYEOV2*	0.023	1.3 (1.0 – 1.6)	0.037	1.1 (1.0 – 1.3)
17	c.601+134G>A	rs9899142	*HOXB13*	0.031	0.7 (0.5 – 1.0)	0.665	1.0 (0.9–1.2)
17	c.1350T>G, p.Tyr450Ter	rs118004742	*EFCAB13*	0.048	1.8 (1.0 – 3.1)	0.637	1.1 (0.8–1.6)
17	c.*2215_*2216insT	rs142044482	*ZNF652*	0.087	1.9 (0.9–3.8)	0.009	0.4 (0.2 – 0.8)
17	g.35766564delA	rs140611363	*ACACA*	0.421	0.9 (0.7–1.1)	0.032	0.9 (0.7 – 1.0)
17	g.35766475A>G	rs72828246	*ACACA*	0.459	0.9 (0.7–1.2)	0.044	0.9 (0.8 – 1.0)
2	g.241075809C>T	rs13406410	*MYEOV2*	0.817	1.0 (0.8–1.3)	0.006	1.2 (1.1 – 1.4)
2	c.2361A>G, p.Thr787	rs61752234	*HDAC4*	0.823	1.1 (0.7–1.6)	0.008	0.7 (0.5 – 0.9)

Abbreviations : Chr, chromosome; OR, Odds ratio

Rs73000144 (c.958C>T, p.Val320Ile) located in *HDAC4* had OR of 14.6 (95% CI 1.5 – 140.2, p-value = 0.018) which was the highest among the statistically significant variants. This variant was very rare, found only in three familial PrCa cases (1.6 %) and in seven sporadic patients (0.6 %) which were all heterozygous for the minor allele. Moreover, the variant was observed only in one of the controls (0.1 %) in heterozygous state.

The rs118004742 nonsense variants (c.1638T>G, p.Tyr546Ter) located in *EFCAB13* was found in total of 15 familial cases of which 12 (6.5 %) were heterozygous and three (1.6%) were homozygous for the minor allele. The variant was found moderately associated with familial PrCa having OR of 1.8 (95% CI 1.0 – 3.1, p-value = 0.048) but was not significant when unselected cases where compared against controls.

Two common non-coding variants in the HOXB3 gene, rs10554930 and rs35384813, had a moderate effect on PrCa risk, with OR values ranging from 1.2

(95 % CI 1.1 – 1.8, p-value = 0.010) to 1.4 (95 % CI 1.1 – 1.8, p-value = 0.013). For the remaining five variants the odds ratios were less than 1 which indicates modulatory role in PrCa.

5.1.2 Novel eQLTs discovered at 2q37 and 17q11.2-q22 loci

The fine-mapping was extended to non-coding variants which might have potentially a regulatory role and thus act as modulators of PrCa risk. In order to discover putative cis-regulatory variants and their target genes, RNA-seq was performed and association between variant data and gene expression was evaluated using two different eQTL analysis approaches. The eQTL analysis was conducted separately for two chromosomal regions and included total of 19 individuals which had targeted sequencing data for 2q37 and 17 individuals which had data for 17q11.2-q22.

In the first approach the eQTL analysis was limited to only those genes that were found to be differentially expressed between the cases and controls and were located within the chromosomal loci being sequenced by targeted sequencing. Variants located within 2MB windows of these genes were then tested for association with the differentially expressed genes. The differential expression analysis resulted in the discovery of all together 8 differentially expressed genes (p-value < 0.05) located in 2q37 and 17q11.2-q22 loci. The following eQTL analysis revealed total of 272 candidate eQTLs. Of all the candidate eQTL variants, the strongest support for regulatory potential was observed for rs11650354. This variant was found to be associated with *TBKBP1* expression, which according to RegulomeDB, has been confirmed by a previous study. Rs12620966, which was associated with *AGAP1* expression in chromosome 2, was considered to have the second highest regulatory potential according to RegulomeDB. This variant overlaps several known TF-binding sites discovered by ChiP-seq studies as well as position weight matrices and TF-footprints discovered by DNaseI footprinting studies.

In the alternative cis-eQTL approach the analysis was limited to 34 known PRCA associated variants obtained from iCOGS dataset which were located within the 2q37 and 17q11.2-q22 loci. The alternative cis-eQTL approach identified only one PrCa-associated candidate eQTL on chromosome 2 and 36 candidate eQTLs on chromosome 17. The strongest evidence of regulatory potential was found for rs4796751 and rs4796616, which are located in chromosome 17. These variants were found to be associated with *DHX58*, *MLX* and *JUP* genes and according to the RegulomeDB both have been previously reported as eQTL variants associated with

MGC20781 and *NT5C3L29* genes. Moreover, they overlap with open chromatin regions in several cell lines.

Furthermore, two chromosome 17 variants, rs4793943 and rs16941107 were found to be eQTL variants by the modified cis-eQTL approach. These variants were found to regulate the expression of *ZNF652* and *ARL17B* genes, respectively, and according to RegulomeDB they overlap with open chromatin regions and TFBS in several cell lines.

5.2 Novel HBOC associated candidate genes and variants (2)

5.2.1 Identifying DNA-repair variants associated with predisposition to breast cancer

In this study, whole exome sequencing was performed for 37 individuals from 13 high-risk BRCA1/2-negative families. This cohort comprised of 23 female breast or breast and ovarian cancer patients, one male BC patient and 13 healthy relatives. The total number of discovered variants in the cohort was 736,963 and further filtering steps focusing on the DNA-repair pathway reduced the number of initial candidate variants to 98.

Eighteen of these initial candidate variants were selected for further validation and testing for association to breast cancer. In total 129 HBOC cases and 989 healthy female controls were genotyped. The results are shown in Table 4. Five of the validated variants including rs1801673, rs4645959, rs2227580, rs2308957 and *RRM2B* c.211dupC were more frequent in female HBOC cases compared to the controls. The odds ratios of these variants ranged from 1.16 to 2.16 which suggests that these variants could possibly be moderate risk variants. However, none of these variants reached statistical significance. Furthermore, rs80357231 located in *BRCA1* which was detected in two affected females in a single breast cancer family was absent in female HBOC patients and healthy controls implicating that this variant is extremely rare.

Table 4. Genotyping results for candidate variants associated with HBOC.

			Carrier frequency					
			Females		Males			
Variant	dbSNP ID	Gene	HBOC cases	Controls	BC cases	Controls	P-value	OR; 95%CI
c.148C>A, p.P50T	rs184042322	AKT2	2/127	5/280	—	—	1	0.88; 0.17–4.57
c.2572T>C, p.F858L	rs1800056	ATM	1/129	10/975	—	—	1	0.75; 0.10–5.92
c.3161C>G, p.P1054R	rs1800057	ATM	1/129	14/981	—	—	1	0.54; 0.07–4.13
c.4424A>G, p.Y1475C	rs34640941	ATM	0/129	1/278	0/49	0/909	1/1[d]	na
c.5558A>T, p.D1853V	rs1801673	ATM	1/129	5/989	—	—	0.52	1.54; 0.18–13.19
c.3904A>C, p.T1302P	rs80357231	BRCA1	0/128	0/986	—	—	1	na
c.496C>T, p.H166Y	rs181044510	CDKN2A	3/129	7/280	—	—	1	0.93; 0.24–3.62
c.77A>G, p.N26S	rs4645959	MYC	5/129[a]	23/987	—	—	0.14	2.02; 0.81–5.01
c.3353A>C, p.Q1118P	rs149561356	NCOA3	0/129	7/279	—	—	0.1	na
c.43G>T, p.V15L	rs2227580	PLAU	2/129	11/984	—	—	0.6	2.16; 0.30–15.45
c.341G>A, p.G114D	rs2308957	RAD1	5/129	15/464	—	—	0.79	1.16; 0.42–3.22
c.280A>C, p.I94L	rs28903085	RAD50	0/129	0/187	0/49	1/909	1/1[d]	na
c.538G>A, p.G180R	rs7487683	RAD52	4/129	15/269	—	—	0.33	0.55; 0.18–1.67
c.1723G>C, p.E575Q	rs76818213	RBL2	8/129	22/261	—	—	0.55	0.73; 0.32–1.66
c.122C>T, p.S41F	rs149249571	RPA2	0/129	5/467	—	—	0.59	na
c.211dupC, p.R71fs	—	RRM2B	16/128[b]	22/247[c]	—	—	0.31	1.39; 0.73–2.64
c.277G>A, p.D93N	rs201274685	WNT3A	1/129	4/468	—	—	1	0.91; 0.10–8.15
c.337C>T, p.R113C	rs141074983	WNT10A	1/129	10/988	—	—	1	0.77; 0.10–6.00

Abbreviations: BC, breast cancer; CI, confidence interval; HBOC, hereditary breast and/or ovarian cancer; OR, odds ratio
[a] Homozygous in 1/129 of the female HBOC cases
[b] Homozygous in 1/128 of the female HBOC cases
[c] Homozygous in 2/247 of the female HBOC controls
[d] Females/Males

Two variants found in the sequenced male breast cancer patient were screened in a cohort consisted of 49 male breast cancer patients and in a cohort consisted of 909 healthy males. The rs28903085 variant located in *RAD50* which was detected in the male BC patient was not observed among the cohort of male breast cancer cases and was only found in one male control. This suggests that it might be rare cancer susceptibility variant conferring to male BC. The other variant rs34640941 located in *ATM*, found in the exome male BC patient cohort, was not found in the validation cohort. Despite some of the variants having odds ratios higher than one, none of the variants were significantly associated with HBOC according to Fisher's exact test, most likely due to the rare occurrence of these variants.

5.2.2 Identifying candidate variants associated with early onset

In order to identify candidate variants associated with early onset of BC variants occurring only in early onset patients were selected for further analysis. After prioritisation of variants based on their assumed pathogenicity, enrichment analysis

was conducted for the target genes of the remaining variants. The enriched terms were related to cell cycle, proliferation, apoptosis adhesion, DNA response and various signalling pathways. The found variants are shown in Table 5.

Table 5. Candidate variants discovered in early-onset breast cancer patients

Gene	Variant	Number of cases	Pathways
AKAP13	c.571G>A, p.(G191R)	1	G Protein signalling
AKAP8	c.1513A>G, p.(N505D)	1	G Protein signalling
APEX1	c.190A>G, p.(I64V)a	1	TSH signalling, base excision repair
BIRC6	c.2675A>G, p.(E892G)	1	Ubiquitin-mediated proteolysis, apoptosis modulation and signalling
BNIPL	c.33dupA, p.(T11fs)a	1	Interacts with BCL2, promotes cell death
BRCA1	c.3155C>T, p.(P1052L)	1	Ubiquitin-mediated proteolysis, DNA damage response
CDC45	c.326A>G, p.(E109G)	1	DNA replication, cell cycle
CDKN2B	c.56C>A, p.(A19D)	1	Cell cycle, TGF beta signalling, pathways in cancer
CHEK2	c.470T>C, p.(I157T)	1	DNA damage response, p53 signalling, cell cycle
CINP	c.159C>G, p.(N53K)	2	DNA replication, checkpoint signalling
COL11A2	c.32T>A, p.(L11H)	1	ECM-receptor interaction, focal adhesion
COL4A6	c.3481A>G, p.(I1161V)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
COL6A2	c.679G>A, p.(D227N)	1	ECM-receptor interaction, focal adhesion
DENND2D	c.46C>T, p.(R16*)	1	Promotes the exchange of GDP to GTP
DHH	c.25C>G, p.(P9A)	1	Hedgehog signalling
DTX4	c.1243C>T, p.(R415C)	1	Notch signalling
EDN3	c.560dupA, p.(E187fs)a	1	Variety of cellular roles including proliferation, migration, differentiation
EFCAB13	c.1009A>T, p.(K337*)	1	Calcium ion binding
EXO1	c.836A>G, p.(N279S)	1	Mismatch repair
FANCD2	c.2702G>T, p.(G901V)	1	DNA damage response
FBXW8	c.1409C>T, p.(T470M)	1	Ubiquitin-mediated proteolysis
FOCAD	c.5047G>A, p.(A1683T)	2	Tumour suppressor in glioma and colorectal cancer
LAMA1	c.2186G>A, p.(R729H)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMA5	c.5035C>T, p.(R1679W)	2	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMA5	c.3062C>T, p.(A1021V)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMA5	c.7367G>A, p.(R2456H)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMA5	c.6413G>T, p.(S2138I)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMB1	c.2869G>A, p.(D957N)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMB2	c.1306G>A, p.(G436S)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LAMC3	c.1687C>T, p.(R563W)	1	ECM-receptor interaction, pathways in cancer, focal adhesion
LIG1	c.841G>A, p.(V281M)	1	DNA replication, mismatch repair, base and nucleotide excision repair
LRP2	c.6850A>G, p.(T2284A)	1	Hedgehog signalling
LRP2	c.5107C>T, p.(P1703S)	1	Hedgehog signalling
MAD1L1	c.175C>T, p.(R59C)	1	Cell cycle, progesterone-mediated oocyte maturation
MAGEF1	c.52dupG, p.(E18fs)a	1	Enhancer of ubiquitin ligase activity
MAP3K4	c.2717A>C, p.(H906P)	1	DNA damage response, MAPK signalling
MBD4	c.1073T>C, p.(I358T)	1	Base excision repair
NEIL3	c.516G>C, p.(Q172H)	1	Base excision repair
NLRP4	c.1912G>A, p.(G638R)	1	NOD signalling
NUMBL	c.1347T>G, p.(F449L)	1	Notch signalling
PLD1	c.1192C>T, p.(R398C)	2	Glycerophospholipid metabolism, pathways in cancer
PRDM1	c.1739C>T, p.(P580L)	1	NOD signalling
RASGRP3	c.844G>A, p.(G282S)	1	Integrated cancer, MAPK signalling
RBL2	c.98A>C, p.(D33A)	2	DNA damage response, TGF beta signalling, cell cycle
RBL2	c.100G>C, p.(A34P)	2	DNA damage response, TGF beta signalling, cell cycle
RBL2	c.179A>G, p.(E60G)	1	DNA damage response, TGF beta signalling, cell cycle
RET	c.2876G>A, p.(R959Q)	1	Pathways in cancer
RICTOR	c.3221A>G, p.(D1074G)	1	TOR signalling, mTOR signalling

Table 5 .Continued

Gene	Variant	Number of cases	Pathways
S1PR5	c.953T>A, p.(L318Q)	2	Signal transduction of S1P receptor
SOX17	c.83G>T, p.(G28V)	1	Wnt signalling
TAB3	c.743C>T, p.(T248M)	1	NOD-like receptor signalling
TICRR	c.1993C>T, p.(R665*)	1	DNA replication
TNC	c.1642G>A, p.(V548M)	1	ECM-receptor interaction, focal adhesion
TNC	c.2977G>C, p.(V993L)	1	ECM-receptor interaction, focal adhesion
UBE2Q1	c.727A>C, p.(N243H)	1	Ubiquitin-mediated proteolysis
UBE3A	c.532G>A, p.(A178T)	1	Ubiquitin-mediated proteolysis

5.3 The effects of BMP4 treatment on transcriptional profiles and chromatin landscape of breast cancer cells (3)

5.3.1 Differential expression and GO enrichment analysis

The effects of BMP4 stimulation on transcriptional regulation and the chromatin landscape was studied using RNA-seq and DNase-seq respectively in MDA-MB-231 and T-47D cell lines. Differential expression analysis between the vehicle treated (unstimulated) condition and BMP4 treated condition yielded 91 differentially expressed genes in MDA-MB-231 cells of which 59 were upregulated and 33 were downregulated. In T-47D, 203 DEGs were found of which 160 were upregulated and 43 were downregulated. Ten of these DEGs were shared by the two cell lines. In order to further investigate the different responses to BMP4 stimulation, GO enrichment analysis was conducted for the sets of differentially expressed genes which were unique to MDA-MB-231 and T-47D. The top enriched categories related to the DEGs found in MDA-MB-231 were related to cell motility and migration whereas the top enriched categories related to the DEGs discovered in T-47D were related to organ development and morphogenesis.

5.3.2 Exploring the temporal patterns of differentially expressed genes in multiple breast cancer cell lines

Based on their known association to cancer and sufficient expression levels observed in RNA-seq 15 DEGs in total were selected for validation in MDA-MB-231, T-47D and five additional breast cancer cell lines at three different time points (3h, 6h, 24h) after stimulation with BMP4. Five of the DEGs (*ATOH8*, *ID2*, *SKIL*, *SMAD6* and

SMAD9) which were observed to be differentially expressed in both MDA-MB-231 and T-47D based on the results obtained from RNA-seq at 3h after stimulation were consistently found to be differentially expressed based on the qPCR in both cell lines as well as the additional cell lines with the exception of MDA-MB-436 in which no significant change in the expression was found. The other 10 DEGs which, based on RNA-Seq, were only differentially expressed in MDA-MB-231 or T-47D, showed more varying transcriptional profiles across the additional cell lines supporting the fact that the response to *BMP4* is more cell type specific.

5.3.3 Alteration in chromatin landscapes of T-47D and MDA-MB-231 after BMP4 stimulation

In MDA-MB-231, 89,830 DNase hypersensitive sites were found in the vehicle-treated sample whereas 97,349 DHS sites were found in the BMP4-treated sample. In T-47D vehicle and BMP4-treated samples the corresponding number of DHS sites were 68,000 and 73,881 respectively. In MDA-MB-231 the total number of DHS sites after merging the overlapping sites across the vehicle and BMP4-treated samples was 106,154 whereas in T-47D the corresponding number of sites was 110,028. In MDA-MB-231 the percentage of shared DHS sites between the two conditions was 75 %. Correspondingly, only 27 % of the DHS sites were shared in T-47D. When comparing the distribution of the DHS sites to different genomic features no significant differences were found between the two cell lines in the vehicle condition. Moreover, in both cell lines BMP4-treatment seemed to increase the proportion of DHS sites located in intronic as well as intergenic sites ,while in other genomic sites the fraction of DHS sites were decreased.

To verify that the effects of stimulation on the chromatin level are consistent with the findings on the transcriptomic level, GO enrichment analysis was conducted for differential DHS sites. The analysis of differential DHS sites in MDA-MB-231 resulted in discovery of enriched categories related to cell motility whereas for T-47D the corresponding analysis highlighted categories related to organ morphogenesis which are consistent with the findings on transcriptomic level. Because of the known relationship between the chromatin structure of TSS and gene expression, coverages of TSS sites of the differentially expressed genes were compared between the vehicle- and BMP4-treated conditions. However, no significant change in the openness of the chromatin was found in any TSS regions of the differentially expressed genes suggesting the BMP4 mediated transcriptional

activation is driven mainly by changes occurring in other gene regulatory regions such as enhancers.

5.3.4 Identified transcription factors involved in BMP4 target gene regulation

To uncover transcriptional regulators involved in the BMP4 response, the promoters of the upregulated genes were scanned using PWMs of 401 TFs. For both cell lines TFs which had enrichment of binding sites in the promoters of the upregulated genes were determined. The TFs which were not expressed in the corresponding cell line were omitted from the list of putative transcriptional regulators. Table 6 shows the 15 most enriched TFs in MDA-MB-231 and T-47D.

Table 6. Top 15 TFs with enriched binding sites in promoters of upregulated genes

TF name	Cell line	Ratio of enrichment	Mean read count
MYBL2	MB-MDA-231	1.92	2197
BACH1	MB-MDA-231	1.81	531
MYC	MB-MDA-231	1.74	3044
MAFK	MB-MDA-231	1.7	688
RELA	MB-MDA-231	1.63	1398
PPARA	MB-MDA-231	1.54	185
NFIA/B/C/X[a]	MB-MDA-231	1.51	[b]
NFIL3	MB-MDA-231	1.48	474
FOXA2	MB-MDA-231	1.47	434
REL	MB-MDA-231	1.47	69
ZFHX3	MB-MDA-231	1.47	66
RXRB	MB-MDA-231	1.47	1015
SMARCC1	MB-MDA-231	1.46	1478
ETV5	MB-MDA-231	1.45	641
NR3C1	MB-MDA-231	1.44	1087
MBD2	T-47D	2.55	571
TFAP2A	T-47D	1.87	941
E4F1	T-47D	1.73	310
SP1	T-47D	1.59	838
CUX1	T-47D	1.5	141
E2F2	T-47D	1.47	215
AHR	T-47D	1.47	791
SP2	T-47D	1.43	672
CREB1	T-47D	1.42	177
CBFB	T-47D	1.42	457
ZIC2	T-47D	1.41	118
ZFX	T-47D	1.37	287
HIF1A	T-47D	1.34	1847
E2F3	T-47D	1.33	322
XBP1	T-47D	1.31	22744

[a] NFIA + NFIB + NFIC + NFIX_f2
[b] Read count range (51, 148, 748, 444), respectively

Three of the enriched TFs (*MBD2*, *CBFB* and *HIF1A*) were selected for further experimental investigation along with SMAD4 which was used as positive control. The four TFs were silenced in both cell lines followed by stimulation with BMP4 or vehicle to experimentally assess the role of these TFs in BMP4 mediated gene regulation. The effects of silencing were then evaluated on the gene expression of the same 15 upregulated genes, which were previously validated with qPCR. As expected SMAD4 silencing resulted in reversal of the BMP4 mediated change of expression in all of the tested genes. Furthermore, silencing *MBD2* was also found to have a similar effect on most of the genes in both cell lines. In contrast, *HIF1A* silencing lead to upregulation of the target genes in MDA-MB-231 in combination with BMP4 stimulation, whereas in T-47D the effect was either opposite or no change in the expression of the target genes. Silencing of *CBFB* lead to abrogation of the BMP4-mediated induction in most of the target genes in T-47D. However, in MDA-MB-231 the downregulation of *CBFB* had variable effects on the target genes response to BMP4.

To further assess the co-regulatory role of *MBD2*, *CBFB* and *HIF1A* with *SMAD4* the co-occurrence of the binding sites of these TFs and four known *SMAD4* motifs in the promoters of upregulated genes was computationally evaluated. As a result several GC-rich motifs were found to be significantly co-occurring with *MBD2* (CGCC, GCCGnCGC and GGCGCC; p-value < 0.001) when the whole set of promoters was used as a background set. For the other two transcription factors none of the consensus motifs were significantly co-occurring in the upregulated promoters.

5.4 Identifying DNA-repair variants associated with aggressive PRCA (4)

A total of 122 lethal PrCa cases of Finnish and Swedish origin and 70 cases unselected for aggressiveness of the disease were screened for germline variants in DNA-repair genes using WES. Due to low overall sequencing depth 10 unselected cases were omitted from the final analysis after sample QC.

Following the GATK best practices workflow a total of 22,850,167 variants across the unselected and lethal cases. After variant filtering and prioritization 31 potentially damaging variants, which were distributed across 17 DNA-repair genes, were discovered among the cases (Table 7). In the control population (n = 9,499) a total of 157 potentially pathogenic variants were discovered in these 17 genes of

which 137 were unique to the control population, giving a total of 168 potentially damaging variants. Of the 168 potentially damaging variants, 47 were classified as Tier 1 variants (predicted deleterious) and 121 as Tier 2 variants (predicted damaging).

Table 7. Predicted damaging mutations discovered in lethal prostate cancer cases

Gene	dbSNP ID	Type	Ref. allele	Alt. allele	Protein change	ClinVar	CADD/REVEL	MAF ExAC	Tier
ATM	rs758081262	stopgain	C	T	Q852X	5	35/-	2,47E−05	1
ATM	rs761486324	frameshift insertion	-	TG	H1082fs	-	-/-	-	1
ATM	rs767099464	frameshift deletion	C	-	H1083fs	-	-/-	-	1
ATM	rs769142993	missense	G	C	A2524P	4	31/0.89	2,48E−05	2
ATM	-	frameshift deletion	AGTAG	-	S2611fs	-	-/-	-	1
ATM	rs753961188	frameshift insertion	-	T	L2885fs	5,4	-/-	4,17E−05	1
ATM	rs376676328	missense	A	G	R2912G	3	29/0.88	3,00E−04	2
BRCA1	rs41293459	missense	C	T	R1699Q	5,4,3	35/0.79	2,49E−05	2
CHEK2	rs555607708	frameshift deletion	G	-	T367fs	5	-/-	1,80E−03	1
CHEK2	rs730881700	frameshift insertion	-	T	E457fs	5,4	-/-	5,02E−05	1
CHEK2	rs137853007	missense	G	A	R145W	5,4	33/0.81	3,30E−05	2
CHEK2	rs28909982	missense	T	C	R117G	5,4	27/0.93	1,00E−04	2
ERCC3	rs753182861	frameshift deletion	T	-	Q586fs	-	-/-	2,00E−04	1
ERCC3	rs145267069	missense	A	G	F297S	-	30/0.82	2,47E−05	2
FAN1	rs778927800	missense	G	A	R749Q	-	34/0.89	8,25E−06	2
FANCM	rs147021911	stopgain	C	T	Q1701X	4	35/0.12	1,30E−03	1
HLTF	rs184046773	missense	C	T	G1886A	-	33/0.81	2,00E−04	2
MRE11A	rs372000848	missense	G	A	R305W	4,3	33/0.85	4,96E−05	2
MUTYH	rs34126013	missense	G	A	R238W	5,4	33/0.79	9,20E−05	2
NEIL1	rs5745906	missense	G	A	G169D	-	27/0.86	1,30E−03	2
NTHL1	rs150766139	stopgain	G	A	Q90X	5,3	35/-	1,50E−03	1
POLG	rs113994097	missense	C	G	W748S	5,3	33/0.91	8,00E−04	2
POLG	rs113994096	missense	G	A	P587L	5,3	28/0.80	1,70E−03	2
POLG	rs121918052	missense	C	G	Q497H	5,3	26/0.71	2,00E−04	2
POLG	rs761584617	missense	G	A	A1115V	-	23/0.80	2,47E−05	2
POLL	rs139871590	missense	C	T	G356S	-	34/0.83	1,00E−03	2
RAD18	rs138830303	stopgain	T	A	K197X	-	36/-	1,00E−04	1
RECQL	rs149937760	missense	C	T	C414Y	-	33/0.84	2,00E−04	2
RECQL5	rs768705080	missense	T	G	Y362S	-	32/0.76	8,24E−06	2
TP53	rs876660754	missense	C	T	V173M	5,4	28/0.89	-	2
TP53	rs779000871	missense	G	A	T170M	3	24/0.87	8,24E−05	2

Abbreviations : Ref.allele, reference allele; Alt.allele, alternative allele; MAF ExAC, Minor allele frequency in ExAC database

In lethal cases, 12.3% carried a potentially damaging Tier 1 variant in a DNA-repair gene. No potentially damaging Tier 1 variants were found in the unselected cases which makes the carrier rate significantly higher in lethal cases (p-value = 0.003). The observed carrier rates of potentially damaging Tier 1 variants were also significantly higher in comparison to the Finnish (p-value = 0.040) and Swedish (p-value < 0.001) control cohorts which had carrier rates of 5.4% and 1,6% respectively. No significant difference in the Tier 1 variant carrier rates were observed between Swedish and Finnish lethal cases (p-value = 0.781). Among the 17 DNA-repair genes

CHEK2 and *ATM* carried the highest number of potentially damaging Tier 1 variants in the lethal cases having carrier rates of 4.1% and 3.3% respectively. The carrier rates for potentially damaging Tier 1 variants are shown in Table 8**.**

Similarly to Tier 1 variants, the carrier rates of the potentially damaging Tier 2 variants were higher in lethal cases compared to unselected cases having carrier rates of 13.1% and 5.0% respectively. However, the difference was not statistically significant (p-value = 0.123). When compared to Swedish controls, the carrier rate was significantly higher (6.8%, p-value = 0.011). In contrast, when compared against the Finnish controls the carrier rate was not significantly higher (9.0%, p-value = 0.148). When comparing potentially damaging Tier 2 variant carrier rates between the Swedish and Finnish lethal cases the difference was not found to be significant (p-value = 0.102). Among the 17 DNA-repair genes *POLG* carried the highest number of potentially damaging Tier 2 variants in the lethal cases having carrier rate 4.1%. The carrier rates for potentially damaging Tier 2 variants are shown in Table 8.

Table 8. Carrier rates of mutations in lethal PrCa, unselected cases and population controls

	Lethal PrCa	Unselected PrCA	p-value	Exome FIN	p-value	Swedish Controls	p-value
Tier 1							
ERCC3, n (%)	1 (0.82)	0	1.000	0	0.036	3 (0.05)	0.075
RAD18, n (%)	1 (0.82)	0	1.000	0	0.036	0	0.019
ATM, n (%)	4 (3.28)	0	0.304	4 (0.12)	< 0.001	10 (0.16)	< 0.001
FANCM, n (%)	2 (1.64)	0	1.000	89 (2.69)	0.772	44 (0.71)	0.223
NTHL1, n (%)	2 (1.64)	0	1.000	24 (0.73)	0.236	39 (0.63)	0.187
CHEK2, n (%)	5 (4.10)	0	0.173	60 (1.81)	0.080	5 (0.08)	< 0.001
All, n (%)	15 (12.30)	0	0.003	177 (5.35)	0.004	101 (1.63)	< 0.001
Tier 2							
MUTYH, n (%)	0	1 (1.67)	0.330	34 (1.03)	0.633	75 (1.21)	0.406
ERCC3, n (%)	1 (0.82)	1 (1.67)	0.552	5 (0.15)	0.195	4 (0.06)	0.093
HLTF, n (%)	1 (0.82)	0	1.000	20 (0.60)	0.534	9 (0.15)	0.177
POLL, n (%)	1 (0.82)	0	1.000	15 (0.45)	0.441	28 (0.45)	0.433
MRE11A, n (%)	1 (0.82)	0	1.000	0	0.036	0	0.019
ATM, n (%)	2 (1.64)	0	1.000	13 (0.39)	0.098	28 (0.45)	0.114
RECQL, n (%)	1 (0.82)	0	1.000	0	0.036	13 (0.21)	0.239
FAN1, n (%)	1 (0.82)	0	1.000	2 (0.06)	0.103	16 (0.26)	0.283
NEIL1, n (%)	1 (0.82)	0	1.000	3 (0.09)	0.135	16 (0.26)	0.283
POLG, n (%)	5 (4.10)	0	0.173	197 (5.96)	0.555	190 (3.07)	0.429
TP53, n (%)	2 (1.64)	0	1.000	3 (0.09)	0.012	7 (0.11)	0.012
BRCA1, n (%)	1 (0.82)	0	1.000	2 (0.06)	0.103	5 (0.08)	0.111
RECQL5, n (%)	1 (0.82)	0	1.000	3 (0.09)	0.135	1 (0.02)	0.038
CHEK2, n (%)	1 (0.82)	1 (1.67)	0.552	2 (0.06)	0.103	28 (0.45)	0.433
All, n (%)	16 (13.11)	3 (5.00)	0.123	299 (9.04)	0.148	420 (6.78)	0.011

6 DISCUSSION

6.1 Development of the framework for variant analysis for studying cancer genetics

Since the discovery of the hereditary component in cancer much effort has been spent in trying to identify genomic loci and variants which contribute to increased cancer risk. Subsequently linkage and GWAS studies have identified numerous loci associated with cancer susceptibility. However, the variants discovered by these studies do not fully explain the observed patterns of increased cancer risk as only relatively common variants have been evaluated.

Next-generation sequencing technology has made it possible to screen the complete genomes of individuals, which have led to the discovery of rare and even so-called private variants occurring only in a single family. While the novel technology has opened up new possibilities in cancer research, it has also brought challenges, which need to be addressed in order to leverage the vast amount of data provided by the technology.

One of the most important steps of the data analysis is mapping the reads back to the reference genome, which is computationally challenging. Because of the sheer amount of data, the algorithms need to maximise the speed in order to make the data analysis feasible often at the expense of accuracy; which can hamper the downstream analysis. Nonetheless, BWA and Bowtie2, which were utilised by the developed framework, have proven to perform well when reads are sufficiently long and data quality is good. (Lee et al 2018; Thankaswamy-Kosalai, Sen and Nookaew 2017).

Accuracy of variant calling is dependent more on the variant calling method used rather than the read mapping algorithm. The early variant calling methods such as Samtools, which was utilised in Study 1, is prone to biases produced by sequencing machines when the phred quality scores are determined. This has an effect on the reliability of the variant calling results and because of this fact the methodology utilised in Study 2 omits the use of base quality scores. However, ignoring the base quality scores will lead to compromised quality of variant calling results without careful manual inspection of each variant call; this was observed in study 2 with a high number of false positive variants discovered. However, manual inspection is

not feasible for large sets of variants. Furthermore, the sequencing technology as well as the read alignment is prone to other types of error, which need to be taken into account during variant calling. For this reason, in study 4 the GATK variant calling pipeline was used which is currently the de facto golden standard in variant analysis. Because GATK is able to apply base quality recalibration, it can make use of the phred quality scores as well as other advantages. When calling for germline variants the variant calling is done first by building all possible haplotypes using a local realignment procedure, which allows correction of some of the mistakes made by the read alignment algorithms. In addition, GATK can utilise the whole cohort being studied for more sensitive variant calling and accurate genotyping. Finally, GATK can utilise machine learning to give a better estimate of how likely the variants are true positive by using various quality metrics and sets of known variants.

While standards for the alignment, post-processing and variant calling have been established in cancer research, the process of identifying variants that are likely to be associated with cancer predisposition remains a bottleneck for the analysis. Because of this, a high proportion of the time and effort of developing this framework was put into this step. When assessing the pathogenicity of variants typically common variants are not of interest with the exception of GWAS studies, which aim for identifying low penetrant variants. Filtering out common variants in the population is dependent on the use of databases including allele frequencies for the population of interest. Since the amount of available genome wide studies was very low during study 1 the source of population level variant allele frequency data was limited to dbSNP, which at that time did not included data from the Finnish population. During studies 2 and 4 the amount of available population allele frequency data had grown to also include individuals from the Finnish population. In study 2 the Sisu database was used which became later part of the ExAC database that was eventually used in study 4. Having a true control cohort matching the cases being studied, made it possible to estimate allele frequencies more accurately and thus avoid selecting variants, which are relatively common in the population being studied.

To further assess the pathogenicity, the variants were prioritised based on their effect on the protein level. In studies 1, 2 and 4 frameshifts and stop gains as well as splice site altering variants were considered most likely to be pathogenic. However, it should be noted that frameshift variants as well as stop gain variants can be tolerated if they occur either near 3' or 5' ends of the transcript due to introduction of novel start codon or alternatively leaving all the functionally important sequence unaffected. Therefore, in study 4 the criteria for frameshifts and stop gain variants

was refined by requiring that a variant must either occur upstream or within a known protein domain to be considered as pathogenic.

To assess the pathogenicity of missense variants several pathogenicity predictor were applied during the development of the framework. In study 1 three pathogenicity predictors were utilised including Mutation Taster, PonP and PolyPhen2. All of these methods are able to take into account several features that are important for the assessment of pathogenicity. Indeed the analysis framework, which was applied, succeeded to identify variants, which were shown to be associated with prostate cancer when larger cohorts were genotyped. However, based on the validation, some of the variants predicted to be benign by all methods were also found to be associated with the increased risk of PrCa. This suggests that these methods were not adequately sensitive even when used in combination. Therefore, in study 2, a broader selection of pathogenicity predictors were chosen for increase the sensitivity of the analysis. During the time of study 4, CADD had become widely used tool in cancer genetics. For that reason, this tool was selected as a first-line predictor for filtering variants. However, based on various benchmarking studies CADD does not provide adequate specificity and therefore REVEL was later used to apply more stringent pathogenicity filtering. REVEL was chosen because it has been designed specifically to evaluated pathogenicity of rare variants and has been reported as among the best performing methods (Alirezaie et al. 2018; Ghosh, Oak, and Plon 2017; Li et al. 2018)

Even if a variant is pathogenic it might not necessarily be associated with the phenotype being studied. For this reason, biological domain knowledge is required to further assess the variant possible association to the phenotype. In study 1 the regions investigated were relatively small and therefore did not cover many genes which therefore required more detailed literature research on genes for efficient prioritisation. Contrary to study 1, in studies 2 and 4 the whole exome was analysed. Inclusion of all known prostate cancer related genes in this case would have led to significantly more extensive candidate variant lists for validation. For this reason focus was put in to DNA-repair variants which have been highlighted by previous studies to be involved in susceptibility to breast cancer as well as aggressive prostate cancer. During studies, 1 and 2 there was very little public data about the relationship of variants and phenotype available in databases. However, at the time of study 4 ClinVar had become a popular resource for interpreting the effects of variants. Therefore, ClinVar became an essential part of the developed framework and complements the use of gene specific databases in the evaluation of pathogenicity.

6.2 Extending the framework for integrative analysis of different NGS applications

In the past two decades, a large number of genome-wide studies focusing on different aspects of molecular biology such as genetic variation, transcriptomics and epigenetics have been carried out. While offering valuable information by themselves, the integration of different data types from the same sample or a sample of similar characteristics such as a cell line will provide more insight into the disease of study.

In study 1, in order to assess non-coding variants, available RNA-seq data was used to evaluate the regulatory potential of the variants using the eQTL approach. Because of the small cohort size in the analysis it is likely that the set of discovered eQTLs have a high proportion of false positives. Therefore, two approaches for pre-filtering either variants or potential target genes was made prior to conducting the eQTL analysis. In the first approach, the strategy was to select only genes that were found to be differentially expressed between the cases and controls. The rationale behind this procedure was that genes which have different expression profiles in cases compared to the controls are most likely to be associated with the pathological processes associated with the development of the disease. Therefore, variants, which alter the expression of these genes, could have modulatory role in cancer predisposition. The second approach involved selection of variants that are known to be associated with PrCa susceptibility. Therefore, instead of assessing the variants possible association to PrCa the purpose of the eQLT analysis was to elucidate the possible mechanism of how these variants could contribute to cancer development.

Even though, the potential number of pairs of variant and target genes being evaluated was drastically reduced by applying the aforementioned approaches, still many candidate eQTLs were discovered. Being aware that many of these eQTLs might be false positive because of the small sample size, additional data obtained from ENCODE studies was utilised as supportive evidence. To this end, RegulomeDB was used as it also includes data from previous eQTL studies. Furthermore, the web interface includes a ranking system based on the supporting evidence, which could be utilised to identify the variants that have the strongest regulatory potential. Finally, several interesting eQTL variants were discovered of which some have been previously described. However, further studies are warranted to show that these variants have a modulatory role in PrCa.

In study 3, due to the lack of replicates, standard statistical methods could not be used. To find potentially differentially expressed genes, predefined thresholds for

log2 fold changes, absolute read counts and their difference between the conditions were used. Even though the analysis lacked statistical rigor, these thresholds were robust enough to identify truly differentially expressed genes according to the validation of qPCR. Furthermore, the results of the enrichment analysis were consistent with known characteristics of these two cell lines when stimulated by BMP4.

The differential DHS sites were evaluated using a previously developed method, which provides an estimate of the relative difference in the strength of the signal of the site being open chromatin between the conditions but does not evaluate the statistical significance of the difference. Although there is no experimental evidence to support these putative differential DHS sites as true findings, the results obtained from enrichment analysis of differential peaks were consistent with the enriched categories found for the unique differentially expressed genes for two cell lines.

In order elucidate the transcriptional regulators involved in the BMP4 response, the promoters of upregulated genes were screened for TFBS and enrichment analysis was then conducted based on the predicted binding sites. In the TFBS prediction the DNase-seq data was leveraged by taking into account only the open chromatin portion of the promoters to reduce the amount of spurious sites. As a result, TFs that were enriched in the promoters of DE genes were found. To shorten the list of likely candidates the knowledge of expression levels of these TFs provided by the RNA-seq was utilised to filter out TFs with low or no expression. This methodology led to discovery of three TFs, which were validated as co-regulators of SMAD, based on a knockout experiment. Figure 3 summarises the final framework developed as a result of the four studies.

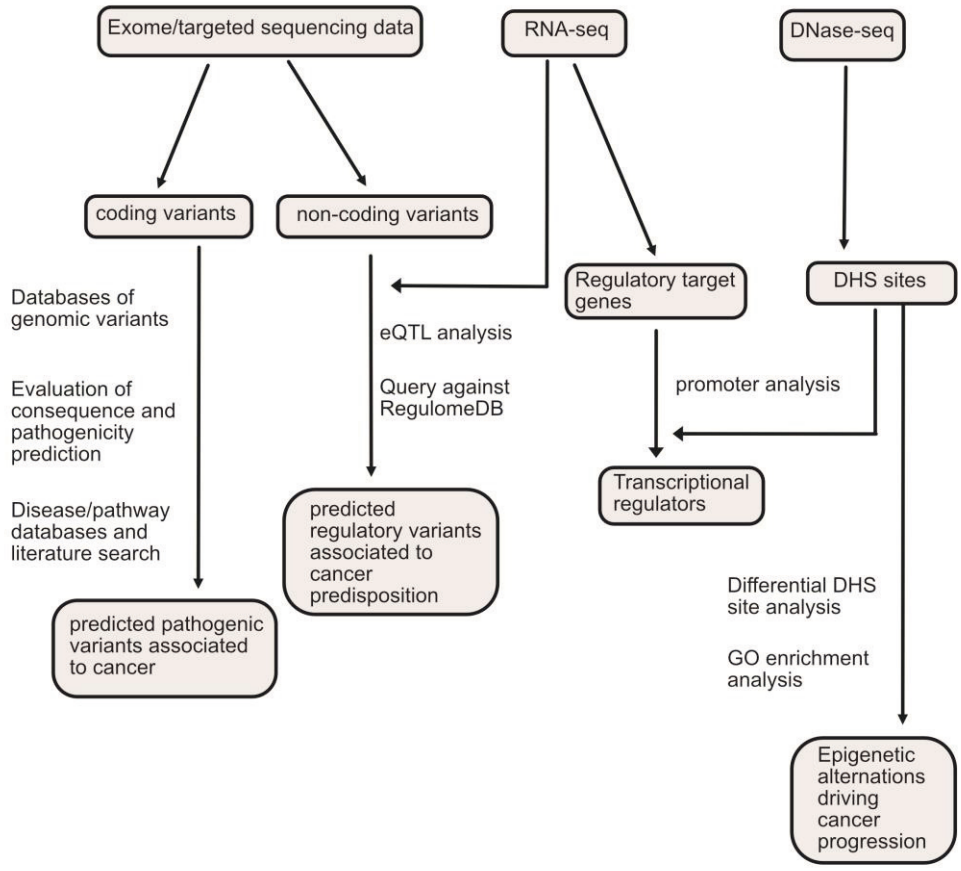

Figure 3. Illustration of the developed framework.

6.3 Challenges and limitations of the study

In study 1 several low to moderate risk variants were found to be associated with PrCa. However, due to low coverage the known G84E *HOXB13* variant could not be detected in the studied cohort although it is relatively common in the Finnish population (Laitinen et al. 2013). This raises the question whether other causal variants could also have been missed. Moreover, the availability of population allele frequency data and tools for assessing the consequences of the variants were very limited at the time. It is likely that due to these limitations potentially pathogenic

variants were left out from validation. This is supported by the fact that some of the predicted neutral variants were found to be associated with increase the risk of PrCa.

Study 2 resulted in the discovery of several clinically interesting DNA-repair variants associated with breast and ovarian cancer. However, their effect size was low or moderate. In addition, very rare variants were found which might confer to higher risk but the clinical significance could not be assessed because of low statistical power. Furthermore, this study concentrated heavily on DNA-repair variants, which is justified by the well-established relationship between DNA-repair variants and cancer predisposition. Still, it is known that in other cancers such as PrCa variants increasing susceptibility are located in genes that are part of a broader range of pathways. Taken together all the shortcomings of the analysis more studies are warranted to identify candidate genes as well as variants contributing hereditary breast and ovarian cancer in the Finnish population.

In the integrative analysis conducted in studies 1 and 3 the major limitations were related to the small sample sizes and lack of biological replicates. In study 1 the number of samples with targeted DNA-seq data and RNA-seq data was less than 20 individuals for the two analysed chromosomal regions. Typical eQTL studies include over 50 individuals and simulation studies suggest that even with one hundred samples the analysis lacks statistical power (Conesa et al. 2016; Huang et al. 2018). Moreover, no correction of multiple testing was performed. Therefore, the results should be considered with caution. Another issue with the eQTL analysis results was that instead of tissue of interest, the prostate, eQTL analysis was conducted only for whole blood namely on lymphocytes. It is known that different tissues have different transcriptional profiles, which makes it uncertain how well the results hold for prostate tissue.

In a study conducted by Schurch et al. 2016, several tools for differential expression analysis were compared in a study in which two conditions were compared having three biological replicates. The conclusions from this study were that none of the methods was able to identify all the true differentially expressed genes. In the light of this study, it is likely that not all the true DEGs could be detected in study 3 as no replicates were included in the RNA-seq analysis. Similarly, in the case of DNaseq-seq the lack of replicates likely compromised the detection of dnase hypersensitive regions and the comparison of differential hypersentive regions between the different conditions.

Due to the lack of a specific SMAD motif, the prediction of its binding sites is unreliable. Because of this, the BMP4 responsive genes could be only detected indirectly based on RNA-seq data. Moreover, the discovery of co-regulatory

transcription factors was challenging. Partly this was because TF footprinting analysis could not be performed because of lack of replicates. TF footprinting would have been likely to reveal candidate TFs more reliably than simply screening the promoters for predicted TFBS.

In study 4, only DNA-repair variants were considered when characterising the germline variants similarly to study 2. While many of the recent studies have focused on the DNA-repair pathway there exists also evidence that germline variants associated with aggressiveness of the disease can be found in genes related to other pathways. In recent study Mijuskovic et al. 2018 found that in addition to DNA-repair genes, pathogenic variants were enriched in a set of genes associated with angiogenesis including *ACVRL1* and *LUM* when the mutational burdens of aggressive cases were compared against indolent cases. Therefore, while using biological domain knowledge based on previous studies can be useful for finding relevant pathogenic variants it also can prevent the discovery of novel candidate genes.

In study 4, unselected PrCa cases as well as the gnomAD database were used as control cohorts. Neither of these cohorts really represent the indolent phenotype as the unselected cases contain relatively aggressive cases and gnomAD can be considered to consist mainly of healthy individuals. Even though differences in the carrier rates between the lethal cohort and the unselected PrCa cohort could be found the comparison against truly indolent cases might have led to more statistically stronger results.

6.4 Future prospects

The sequencing technologies as well as tools for analysing the data are rapidly evolving. As more data is being produced, more databases become available and the existing ones are becoming more comprehensive. While these advances will foster the development of the field of biomedical research they will also render the currently existing frameworks for analysing the data obsolete. Therefore, the frameworks need to be updated as new technologies, tools and publicly available data become available.

6.4.1 Developing framework for variant analysis for identifying cancer associated variants

One of the most difficult task when analysing vast amounts variant data is to find the most likely variants to be associated with the studied disease. Especially in the case when rare variants are considered this question becomes important since often the typical case-control design applied in GWAS studies cannot be applied. Due to this fact, the currently developed frameworks often utilise in-silico tools for assessing the overall pathogenicity of the variants as well as databases that relate genes to pathways and diseases.

In the past the judgement of variant being pathogenic has been highly subjective and no standards have existed. However, now when assessing the pathogenicity of variants it has become increasingly common to refer to guidelines set by the American College of Medical Genetics and Genomics and Association of Molecular Pathology (ACMG-AMP) (Richards et al. 2015). Future frameworks are likely to be designed according to these guidelines.

While the ACMG-AMP guidelines provide instructions how to classify variants and based on this classification assess their potential to affect the phenotype, no suggestions are made on the tools, which are most appropriate to be used (Richards et al. 2015). Many of the past studies as well as the framework developed during studies 1,2 and 4 utilised the classification of variants similar to the one outlined by the ACMG-AMP guidelines followed by prioritisation of the variants based on these categories. It should be noted, that there are many tools available, which could be used to further prioritise the variants that fall into these categories. This is essential because not all variants belonging to the same category can be considered equally pathogenic. As an example, splice site acceptor and donor variants, which are commonly considered potentially highly pathogenic, may alter a splice site but not have a serious consequence because they might generate a novel cryptic exon, which can restore the functionality of the protein (Zavolan et al. 2003). Another example of variants that are generally considered highly pathogenic are the stop gain variants introduced by SNVs. However, it has been shown that transcripts stop gain variants can be rescued by alternate transcription starts sites when the SNVs occur in the near the 5' end of the gene (Cohen et al. 2019). Moreover, the transcript can be rescued also when the SNV occurs near the 3' end of the gene as the sequence being affected might not change the proteins characteristics. Currently, there exists already a wide variety of tools, which are designed to assess the characteristics of specific variant types such as frameshifts, stop gains, splice site donor and acceptor variants,

missense variants and non-coding variants (Duzkale et al. 2013). In the future bioinformatics analysis frameworks would substantially benefit by integrating some of these variant type specific prediction tools.

Besides the overall pathogenicity of the variant, it is important to be able to predict whether the presumed pathogenic variants are associated with a disease. Pathway and Gene databases are constantly being expanded as new studies are published which will directly benefit the future frameworks. Furthermore, as large variant databases such gnomAD have become available it is now possible to compare the rates of different types of variants observed in genes. It has been shown that some genes are more intolerant of variants commonly considered as pathogenic such as stop gain variants in comparison to others and are therefore more susceptible. Based on the variable variant counts observed in genes, metrics have been developed for evaluating the gene specific evolutionary constraints (Karczewski et al. 2019). These metrics could be used as complementary information to gene and pathway databases in future frameworks.

Because of the challenges faced when dealing with rare variants statistical methods that attempt to aggregate the variants on gene level called burden tests are becoming increasingly popular (Lee et al. 2014). However, the application of a burden test is often impossible because of lack of suitable control cohorts due to the high sequencing costs of producing them. As a solution, strategies, which can utilise large databases as control populations have been suggested (Guo et al. 2018). Application of this strategy can be utilised for candidate gene discovery along with the above-mentioned methods.

6.4.2 Developing integrative approaches for studying the relationship of variants and gene expression

The eQTL analysis requires sufficient sample sizes to reach statistically significant results. This needs to be taken into account when designing the study. Another challenge is the vast amount of false positive eQTLs that are obtained because of the multiple hypothesis problem. Traditional p-value adjustment methods have been found to be too stringent to be utilised for controlling the false discovery rate as variants do not occur completely independently but are dependent on the LD structure (Huang et al. 2018). The more recently developed tools are able to take into account the LD structure and thus are able to alleviate the issue of too stringent p-value adjustment (Davis et al. 2016; Johnson et al. 2010). Therefore, the use of these

methods lowers the risk of filtering out too many eQTL candidates while still keeping the amount of false positive variants at low levels.

Another approach related to the study design would be to incorporate epigenetic profiling using techniques such as ATAC-seq. This information can be used to prioritise variants that are located in regions of active regulatory sites and detect occurrence of variants in TF-footprints. Together with ENCODE data this would make it possible to predict whether putative eQTL variants occurring within TFBS offering supportive evidence for regulatory role of these variants as transcriptional regulators.

6.4.3 Developing of framework for studying epigenetic data and transcriptional regulators in cancer progression

Recently, ATAC-seq has become more a popular method compared to DNase-seq for characterisation of the chromatin landscape because it is simple to perform and still provides accurate results. Another advantage of ATAC-seq is the possibility to also study clinical samples. This allows the characterisation of the epigenetic profiles of tumour samples, which have been shown to define different subtypes of cancer (Corces et al. 2018). In future studies ATAC-seq and RNA-seq could be used to investigate transcriptomic and epigenetic profiles of a population of tumours. This could reveal how different subtypes arise and what are the key transcriptional regulators driving tumour progression. The epigenetic patterns and transcriptional regulators could be further studied in more controlled manner with cancer cell lines. Similar to study 3, the key regulators could be either knocked out or alternatively pathways involved in cancer progression could be stimulated. At this stage, methods such as ChIP-seq could also be incorporated to study individual transcription factors.

7 CONCLUSIONS

The current study was conducted to develop a bioinformatics framework for analysis of germline variant data which was applied to uncover variants which are associated with hereditary prostate, breast and ovarian cancer. Moreover, the framework was extended to include integrative analysis of NGS data to elucidate the regulatory potential of variants as well as to study epigenetic changes occurring during cancer development. This extended framework was then applied to the discovery of eQTLs which have a potential modulatory role in PrCa predisposition and to study BMP4 response in breast cancer.

1. The bioinformatics framework developed for variant analysis revealed several low to moderate risk variants associated with PrCa and HBOC. In addition, novel and known pathogenic variants associated with PrCa were identified in lethal prostate cancer cases as well as control cohorts, which allowed the comparison of carrier frequencies between the cohorts. It was found that the *CHEK2* and *ATM* rather that *BRCA2* were the genes having the most highest frequencies of presumed pathogenic variants and therefore likely to have more prominent role in predisposition to aggressive PrCa in Finnish and Swedish populations.

2. The eQTL analysis led to discovery of non-coding variants, which are associated with regulation of genes that were found to be differentially expressed between affected individuals and controls. This suggests that they might have a modulatory role in the development of prostate cancer. Moreover, this study revealed additional non-coding eQTL variants, which have been previously shown to be associated with prostate cancer susceptibility. The uncovered relationship between these variants and their potential target genes might explain the mechanism of how they contribute to the development of PrCa.

3. The integrative analysis framework revealed major differences in the chromatin landscape of the two breast cancer cell lines, which likely explains the different BMP4 response. Furthermore, BMP4 response genes key regulatory transcription factors associated with the regulation of the response genes were identified.

REFERENCES

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;(SUPPL.76).

Ahmed N, Bertels K, Al-Ars Z. A comparison of seed-and-extend techniques in modern DNA read alignment algorithms. In: Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016. 2017;1421–8.

Alirezaie, Najmeh, Kristin D. Kernohan, Taila Hartley, Jacek Majewski, and Toby Dylan Hocking. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. American Journal of Human Genetics. 2018;103 (4): 474–83.

Amberger, JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders. Nucleic Acids Research 43 (Database issue): D789–98.

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10): R106.

Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.

Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. 2010. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Annala M. Pypette. 2016. Available online at: https://github.com/annalam/pypette.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015 (526); 68–74.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics. 2011;(12):745–55.

Bansal V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. BMC Bioinformatics. 2017;18.

Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15.

Benafif S, Kote-Jarai Z, Eeles RA. A review of prostate cancer Genome-Wide Association Studies (GWAS). Cancer Epidemiology Biomarkers and Prevention. 2018 (27): 845–57.

Beral V, Bull D, Doll R, Peto R, Reeves G, Skegg D, et al. Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. Lancet. 2001;358(9291):1389–99.

Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018;33(4):690-705.e9.

Birbrair A. Stem cell microenvironments and beyond. In: Advances in Experimental Medicine and Biology. 2017;1–3.

Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biology. 2017 (18).

Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: A feature density estimator for high-throughput sequence tags. Bioinformatics. 2008; 24(21):2537–8.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790–7.

Broad Institute. Picard. 2016. Available online at : http://broadinstitute.github.io/picard/.

Canzar S, Salzberg SL. Short Read Mapping: An Algorithmic Tour. In: Proceedings of the IEEE. 2017;436–58.

Carter, HB, Helfand B, Mamawala M, Wu Y, Landis P, Yu H, Wiley K, et al. Germline Mutations in ATM and BRCA1/2 Are Associated with Grade Reclassification in Men on Active Surveillance for Prostate Cancer. European Urology 2019;75(5):743–49.

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39 (SUPPL. 1).

Chang P, Gohain M, Yen MR, Chen PY. Computational Methods for Assessing Chromatin Hierarchy. Computational and Structural Biotechnology Journal. 2018;(16):43–53.

Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics. 2017;18.

Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. In: Bioinformatics. 2018;i884–90.

Chung JC, Chen SL. Lacer: Accurate Base Quality Score Recalibration For Improving Variant Calling From Next-Generation Sequencing Data In Any Organism. bioRxiv [Internet]. 2017;130732. Available from: https://www.biorxiv.org/content/early/2017/04/27/130732

Cohen S, Kramarski L, Levi S, Deshe N, Ben David O, Arbely E. Nonsense mutation-dependent reinitiation of translation in mammalian cells. Nucleic Acids Res. 2019;47(12):6330–8.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;(17).

Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science (80-). 2018;362(6413).

Cropp CD., Simpson CL, Wahlfors T, Ha N, George A, Jones MS, Harper U, 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. International Journal of Cancer. Journal International Du Cancer 2011;129(10):2400–2407.

Davis JR, Fresard L, Knowles DA, Pala M, Bustamante CD, Battle A, et al. An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. Am J Hum Genet. 2016;98(1):216–24.

Demichelis F, Stanford JL. Genetic predisposition to prostate cancer: Update and future perspectives. Vol. 33, Urologic Oncology: Seminars and Original Investigations. 2015;75–84.

Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–501.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

Duzkale, H., J. Shen, H. McLaughlin, A. Alfares, M. A. Kelly, T. J. Pugh, B. H. Funke, H. L. Rehm, and M. S. Lebo. 2013. A Systematic Approach to Assessing the Clinical Significance of Genetic Variants. Clinical Genetics 84 (5): 453–63.

Eeles RA, Olama AA Al, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013;45(4):385–91.

Eilbeck K, Quinlan A, Yandell M. Settling the score: Variant prioritization and Mendelian disease. Nature Reviews Genetics. 2017;(18):599–612.

Elgin SCR. DNAase I-hypersensitive sites of chromatin. Vol. 27, Cell. 1981. p. 413–5.

Fischer D. The R-package GenomicTools for multifactor dimensionality reduction and the analysis of (exploratory) Quantitative Trait Loci. Comput Methods Programs Biomed. 2017;151:171–7.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2013;41(D1).

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2011;39(SUPPL. 1).

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216–20.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, et al. The UCSC genome browser database: Update 2011. Nucleic Acids Res. 2011;39(SUPPL. 1).

Galas DJ, Schmitz A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5(9):3157–70.

García-García G, Baux D, Faugère V, Moclyn M, Koenig M, Claustres M, et al. Assessment of the latest NGS enrichment capture methods in clinical context. Sci Rep. 2016;6.

Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(90001):258D – 261.

Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18(1).

Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: An accessible system for testing SNV novelty. Bioinformatics. 2011;(27):3216–7.

Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;(17):333–51.

Guo MH, Plummer L, Chan YM, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. Am J Hum Genet. 2018;103(4):522–34.

Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2013;15(6):879–89.

Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011;(144):646–74.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;22(9):1760–74.

Hayden EC. Is the $1,000 genome for real? Nature. 2014. https://doi.org/10.1038/nature.2014.14530.

He, HH, Meir CA, Chen MW, Jordan VC, Brown M, and Liu XS. Differential DNase I Hypersensitivity Reveals Factor-Dependent Chromatin Dynamics. Genome Research 2012;22(6):1015–25.

Hjelmborg JB, Scheike T, Holst K, Skytthe A, Penney KL, Graff RE, et al. The heritability of prostate cancer in the Nordic twin study of cancer. Cancer Epidemiol Biomarkers Prev. 2014;23(11):2303–10.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

Huang QQ, Ritchie SC, Brozynska M, Inouye M. Power, false discovery rate and Winner's Curse in eQTL studies. Nucleic Acids Res. 2018;46(22).

Huo Y, Li S, Liu J, Li X, Luo XJ. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. Nat Commun. 2019;10(1).

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877–85.

Jalali SDM, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques. 2017;62(1):18–30.

Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, et al. Genome-wide detection of DNase hypersensitive sites in single cells and FFPE tissue samples. Nature. 2015;528(7580):142–6.

Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010;11(1).

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009;458(7236):362–6.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv [Internet]. 2019;531210. Available from: https://www.biorxiv.org/content/10.1101/531210v3

Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: Displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840–5.

Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(90001):493D – 496.

Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4).

Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591–4.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.

Kulakovskiy I V., Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. 2013;41(D1).

Kulakovskiy I V., Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46(D1):D252–9.

Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotechnol. 2013;31(7):615–22.

Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11).

Laitinen V. Genetic Risk Factors for Hereditary Prostate Cancer in Finland. 2016. Acta Universitatis Tamperensis 2179. Tampere University Press. Tampere.

Laitinen VH, Wahlfors T, Saaristo L, Rantapero T, Pelttari LM, Kilpivaara O, et al. HOXB13 G84E mutation in Finland: Population-based analysis of prostate, breast, and colorectal cancer risk. Cancer Epidemiol Biomarkers Prev. 2013;22(3):452–60.

Lakeman IMM, Schmidt MK, van Asperen CJ, Devilee P. Breast Cancer Susceptibility—Towards Individualised Risk Prediction. Curr Genet Med Rep. 2019;7(2):124–35.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

Lee C. Genome-wide expression quantitative trait loci analysis using mixed models. Frontiers in Genetics. 2018:(9).

Lee, H, Lee K-W, Lee T, Park D, Chung J, Lee C, Park W-Y, and Son D-S. Performance Evaluation Method for Read Mapping Tool in Clinical Panel Sequencing. Genes & Genomics 2018;40(2):189–97

Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. Vol. 95, American Journal of Human Genetics. 2014;5–23.

Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum Mutat. 2015;36(8):815–22.

Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

Li, J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, and Sun Z. Performance Evaluation of Pathogenicity-Computation Methods for Missense Variants. Nucleic Acids Research 2018;46(15):7793–7804.

Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common genetic variation and breast cancer Risk—Past, present, and future. Cancer Epidemiology Biomarkers and Prevention. 2018;(27):380–94.

Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, et al. Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. PLoS Genet. 2014;10(7).

Lindmark F, Zheng SL, Wiklund F, Bensen J, Bälter KA, Chang B, et al. H6D polymorphism in macrophage-inhibitory cytokine-1 gene associated with prostate cancer. J Natl Cancer Inst. 2004;96(16):1248–54.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12).

Luehr S, Hartmann H, Söding J. The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. Nucleic Acids Res. 2012;40(W1).

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science (80-). 2012;335(6070):823–8.

Madsen JGS, Rauch A, Van Hauwaert EL, Schmidt SF, Winnefeld M, Mandrup S. Integrated analysis of motif activity and gene expression changes of transcription factors. Genome Res. 2018;28(2):243–55.

Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: From SNPs to phenotypes. Trends in Genetics. 2011;(27):72–9.

Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res. 2003;31(20):6016–26.

Maqungo M, Kaur M, Kwofie SK, Radovanovic A, Schaefer U, Schmeier S, et al. DDPC: Dragon database of genes associated with prostate cancer. Nucleic Acids Res. 2011;39(SUPPL. 1).

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28(5):495–501.

Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Briefings in Functional Genomics. 2011;(10):374–86.

Metzker ML. Sequencing technologies the next generation. Nature Reviews Genetics. 2010;(11):31–46.

Mijuskovic M, Saunders EJ, Leongamornlert DA, Wakerell S, Whitmore I, Dadaev T, et al. Rare germline variants in DNA repair genes and the angiogenesis pathway predispose prostate cancer patients to develop metastatic disease. Br J Cancer. 2018;119(1):96–104.

Miyamoto K, Nguyen KT, Allen GE, Jullien J, Kumar D, Otani T, et al. Chromatin Accessibility Impacts Transcriptional Reprogramming in Oocytes. Cell Rep. 2018;24(2):304–11.

Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial risk and heritability of cancer among twins in nordic countries. JAMA - J Am Med Assoc. 2016;315(1):68–76.

Määttä K. Genetic Predisposition to Breast and Ovarian Cancer. 2016. Acta Universitatis Tamperensis 2140. Tampere University Press. Tampere.

Na, RS. Zheng L, Han M, Yu H, Jiang D, Shah S, Ewing CM, et al. Germline Mutations in ATM and BRCA1/2 Distinguish Risk for Lethal and Indolent Prostate Cancer and Are Associated with Early Age at Death. European Urology 2017;71(5):740–47.

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863–74.

Nica AC, Dermitzakis ET. Expression quantitative trait loci: Present and future. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013;(368).

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 1999;(27)29–34.

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. PON-P: Integrated predictor for pathogenicity of missense variants. Hum Mutat. 2012;33(8):1166–74.

Pico AR, Kelder T, Van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: Pathway Editing for the People. PLoS Biology. 2008;(6):1403–7.

Pritchard, CC, Mateo J, Walsh MF, Sarkar ND, Abida W, Beltran H, Garofalo A, et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. The New England Journal of Medicine 2016;375(5):443–53.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

Qi J, Asl HF, Björkegren J, Michoel T. KruX: Matrix-based non-parametric eQTL discovery. BMC Bioinformatics. 2014;15(1).

Quinlan AR. BEDTools: The Swiss-Army tool for genome feature analysis. Curr Protoc Bioinforma. 2014;2014:11.12.1-11.12.34.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol. 2011;12(7).

Reinert K, Langmead B, Weese D, Evers DJ. Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Hum Genet. 2015;16(1):133–51.

Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.

Reuter JA, Spacek D V., Snyder MP. High-Throughput Sequencing Technologies. Molecular Cell. 2015;(58):586–97.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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. 2015;17(5):405–24.

Rivandi M, Martens JWM, Hollestelle A. Elucidating the underlying functional mechanisms of breast cancer susceptibility through post-GWAS analyses. Frontiers in Genetics. 2018:(9).

Robinson PN, Rosario MP, Jager M. Computational Exome and Genome Analysis. Chapman & Hall/CRC Computational biology series. 2017

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, et al. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods. 2006;3(7):511–8.

Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: Advances and future challenges. Nucleic Acids Research. 2014;(42):8845–60.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.

Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6(3).

Schumacher, FR., Olama AAA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, Dadaev T, et al. Association Analyses of More than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci. Nature Genetics 2018;50(7):928–36.

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? Rna. 2016;22(6):839–51.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods. 2010;(7):575–6.

Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. Nature Genetics. 2016;(48):1071–9.

Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–8.

Sherry ST. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.

Sieber KB, Batorsky A, Siebenthall K, Hudkins KL, Vierstra JD, Sullivan S, et al. Integrated functional genomic analysis enables annotation of kidney genome-wide association study loci. J Am Soc Nephrol. 2019;30(3):421–41.

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart - Biological queries made easy. BMC Genomics. 2009;10.

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer. 2018;(18):696–705.

Song L, Crawford GE. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;5(2).

Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature Reviews Genetics. 2019;(20):631–56.

Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 2011;12(10).

Sung MH, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol Cell. 2014;56(2):275–85.

Syrjäkoski K, Hyytinen ER, Kuukasjärvi T, Auvinen A, Kallioniemi OP, Kainu T, et al. Androgen receptor gene alterations in Finnish male breast cancer. Breast Cancer Res Treat. 2003;77(2):167–70.

Syrjäkoski K, Kuukasjärvi T, Waltering K, Haraldsson K, Auvinen A, Borg Å, et al. BRCA2 mutations in 154 Finnish male breast cancer patients. Neoplasia. 2004;6(5):541–5.

Takata, R, Takahashi A, Fujita M, Momozawa Y, Saunders EJ, Yamada H, Maejima K, et al. 12 New Susceptibility Loci for Prostate Cancer Identified by Genome-Wide Association Study in Japanese Population. Nature Communications. 2019;10 (1): 4422.

Thankaswamy-Kosalai, S, Sen P, and Nookaew I.. Evaluation and Assessment of Read-Mapping by Multiple next-Generation Sequencing Aligners Based on Genome-Wide Characteristics. Genomics 2017;109(3-4):186–91.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82.

Valencia AM, Kadoch C. Chromatin regulatory mechanisms and therapeutic opportunities in cancer. Nature Cell Biology. 2019;(21):152–61.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;(SUPL.43).

Venter CJ, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (80-). 2001;291(5507):1304–51.

Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nature Medicine. 2004;(10):789–99.

Vorontsov IE, Kulakovskiy I V., Makeev VJ. Jaccard index based similarity measure to compare transcription factor binding site models. Algorithms Mol Biol. 2013;8(1).

Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16).

Wang X. Next-Generation Sequencing Data Analysis. CRC Press. 2016

Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;(10):57–63.

Warner EW, Yip SM, Chi KN, Wyatt AW. DNA repair defects in prostate cancer: impact for screening, prognostication and treatment. BJU International. 2019;(123):769–76.

Weintraub H, Groudine M. Chromosomal subunits in active genes have an altered conformation. Science (80-). 1976;193(4256):848–56.

Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research. 2018;7:1338.

Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. Genome Medicine. 2014;(6).

Zavolan M, Kondo S, Schönbach C, Adachi J, Hume DA, Arakawa T, et al. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Research. 2003;(13):1290–300.

Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, Jiang X, et al. Genome-Wide Association Study Identifies 32 Novel Breast Cancer Susceptibility Loci from Overall and Subtype-Specific Analyses. Nature Genetics 2020;52(6):572–81.

Zhang Y, Manjunath M, Zhang S, Chasman D, Roy S, Song JS. Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer–associated genetic variant rs4415084. Cancer Res. 2018;78(7):1579–91.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9).

Zhou Q, Su X, Wang A, Xu J, Ning K. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. PLoS One. 2013;8(4).

PUBLICATIONS

PUBLICATION I

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

Laitinen VH, Rantapero T, Fischer D, Vuorinen EM, Tammela TL; PRACTICAL Consortium, Wahlfors T, Schleutker J.

Int J Cancer. 2015. May 15;136(10):2316-27.
doi: 10.1002/ijc.29276.

HHS Public Access

Author manuscript

Int J Cancer. Author manuscript; available in PMC 2016 May 15.

Published in final edited form as:

Int J Cancer. 2015 May 15; 136(10): 2316–2327. doi:10.1002/ijc.29276.

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, FI-20014 University of Turku, Turku, Finland

Abstract

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 SNPs possibly regulating six genes that were differentially expressed between cases and controls. In a modified approach, pre-filtered 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.

Keywords

prostate cancer risk; genetic predisposition; susceptibility loci; 2q37; 17q11.2-q22

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; Johanna.Schleutker@utu.fi.
[§]Full list after Acknowledgements

The authors have declared that no conflicts of interest exist.

Introduction

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 (GWAS).[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 re-sequencing 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 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 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.

Materials 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 re-sequencing 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 re-sequencing of the 2q37 and 17q11.2-q22 regions (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, Inc., Madison, WI, USA) and were sequenced on a Genome Analyzer IIx (Illumina, Inc., San Diego, CA, USA) 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 non-synonymous 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 (MAF) 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.

Validation of predicted PrCa-associated variants with Sequenom

After filtering, 58 variants in 35 target genes (listed in Tables S2–S4) were selected for validation which was performed on germline DNA from 2216 subjects, including 1293 cases and 923 population controls. The majority of the cases (1105 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, Inc., San Diego, CA, USA). 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 non-coding 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 co-segregation 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 re-sequencing step (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, USA) 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, USA). The massively parallel paired-end RNA sequencing was performed at Beijing Genomics Institute (BGI Hong Kong Co., Ltd., Tai Po, Hong Kong) using an Illumina HiSeq2000 sequencing platform (Illumina Inc.).

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 cut-off 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 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 pre-identify 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 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 (Figure 1). Among the 41 predicted pathogenic variants in 2q37, there were 20 missense SNVs, 16 non-coding 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 non-coding 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 (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, 2, 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 non-coding 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 = 3.3×10^{-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 be benign or neutral by all three in silico pathogenicity prediction algorithms (Table S2).

The rs118004742 nonsense mutation (*c.1638T>G, p.Tyr546Ter*) in the *EFCAB13* gene was predicted to be pathogenic by MutationTaster (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; Table S4).

Two common non-coding variants in the *HOXB3* gene, rs10554930 and rs35384813, had a moderate effect on PrCa risk, with OR values ranging from 1.2 to 1.4 (Tables 1 and 2). MutationTaster predicted both of these variants to be pathogenic (Table S2). For five variants, the odds ratios 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 (Figure 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 co-segregation with affection status in 20 families (Tables S6, S7 and 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 (Figure S2A), as did rs79670217 (Figure S2B). Complete segregation of rs118004742 was observed in three families (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 (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.

eQTL mapping results

Differential gene expression analysis revealed three genes (out of 173 tested) located at 2q37 and five genes (out 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 (Table S5). Altogether, 272 candidate regulatory SNPs were identified for six DE genes only (Table S10). A vast majority, 237 candidate SNPs potentially regulate the expression of *AGAP1, SCLY* and *NDUFA10* at 2q37 (Figure 2). The remaining 35 candidate SNPs possibly regulate *TBKBP1, PNPO* and *NAGS* at 17q11.2-q22 (Figure 3). Based on the ENCODE data, the strongest evidence for regulatory potential was found for rs11650354 on chromosome 17, which targets the *TBKBP1* 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 (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 re-sequencing. 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 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 (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 re-sequenced 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 co-expression 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 an extremely 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 co-segregation 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 (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 (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 (Table S2). The variant segregated completely with affection status in three Finnish mutation-positive HPC families and showed evidence for partial co-segregation 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 (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. Co-

segregation of the *ZNF652* variant rs79670217 (Table S7) and G84E was detected in 6/42 (14%) of affected individuals, and among the 31 PrCa patients carrying the *EFCAB13* variant rs118004742 (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 non-coding 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. Post-mortem 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* (Figures 2 and 3, 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 (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 (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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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. The genotyping of variants with Sequenom was performed by the Technology Centre, Institute of Molecular Medicine (FIMM), University of Helsinki, Finland. This work was supported by the Academy of Finland [251074]; The Finnish Cancer Organisations; the Sigrid Juselius Foundation; and the Competitive State Research Financing of the Expert Responsibility Area of Tampere University Hospital [X51003]. The PRACTICAL consortium was supported by the European Commission's Seventh Framework Programme [HEALTH-F2-2009-223175]; Cancer Research UK [C5047/A7357, C1287/A10118, C5047/A3354, C5047/A10692, C16913/A6135]; and The National Institutes of Health [Cancer Post-Cancer GWAS initiative grant No. 1 U19 CA 148537-01].

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

§ The PRACTICAL consortium

Rosalind Eeles[1,2], Doug Easton[3], Kenneth Muir[4], Graham Giles[5,6], Fredrik Wiklund[7], Henrik Grönberg[7], Christopher Haiman[8], Johanna Schleutker[9,10], Maren Weischer[11], Ruth C. Travis[12], David Neal[13], Paul Pharoah[14], Kay-Tee Khaw[15], Janet L. Stanford[16,17], William J. Blot[18], Stephen Thibodeau[19], Christiane Maier[20,21], Adam S. Kibel[22,23], Cezary Cybulski[24], Lisa Cannon-Albright[25], Hermann Brenner[26], Jong Park[27], Radka Kaneva[28], Jyotnsa Batra[29], Manuel R. Teixeira[30], Zsofia Kote-Jarai[1],Ali Amin Al Olama[3], Sara Benlloch[3]

[1] The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey, SM2 5NG, UK,[2] Royal Marsden NHS Foundation Trust, Fulham and Sutton, London and Surrey, UK,[3] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, UK,[4] University of Warwick, Coventry, UK,[5] Cancer Epidemiology Centre, The Cancer Council Victoria, 1 Rathdowne street, Carlton Victoria, Australia,[6] Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Victoria, Australia,[7] Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden,[8] Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, USA,[9] Department of Medical Biochemistry and Genetics, University of Turku, Turku, Finland,[10] BioMediTech, University of Tampere and FimLab Laboratories, Tampere, Finland,[11] Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, Herlev Ringvej 75, DK-2730 Herlev, Denmark,[12] Cancer Epidemiology Unit, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK,[13] Surgical Oncology (Uro-Oncology: S4), University of Cambridge, Box 279, Addenbrooke's Hospital, Hills Road, Cambridge, UK and Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK,[14] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, UK,[15] Cambridge Institute of Public Health, University of Cambridge, Forvie Site, Robinson Way, Cambridge CB2 0SR,[16] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA,[17] Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, USA,[18] International Epidemiology Institute, 1455 Research Blvd., Suite 550, Rockville, MD 20850,[19] Mayo Clinic, Rochester, Minnesota, USA,[20] Department of Urology, University Hospital Ulm, Germany,[21] Institute of Human Genetics University Hospital Ulm, Germany,[22] Brigham and Women's Hospital/Dana-Farber Cancer Institute, 45 Francis Street- ASB II-3, Boston, MA 02115,[23] Washington University, St Louis, Missouri,[24] International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland,[25] Division of Genetic Epidemiology,

Department of Medicine, University of Utah School of Medicine,[26] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg Germany,[27] Division of Cancer Prevention and Control, H. Lee Moffitt Cancer Center, 12902 Magnolia Dr., Tampa, Florida, USA,[28] Molecular Medicine Center and Department of Medical Chemistry and Biochemistry, Medical University - Sofia, 2 Zdrave St, 1431, Sofia, Bulgaria,[29] 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,[30] Department of Genetics, Portuguese Oncology Institute, Porto, Portugal and Biomedical Sciences Institute (ICBAS), Porto University, Porto, Portugal

References

1. Baker SG, Lichtenstein P, Kaprio J, Holm N. Genetic susceptibility to prostate, breast, and colorectal cancer among Nordic twins. Biometrics. 2005; 61:55–63. [PubMed: 15737078]
2. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, Wiley KE, Isaacs SD, Johng D, Wang Y, Bizon C, Yan G, et al. Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med. 2012; 366:141–149. [PubMed: 22236224]
3. Laitinen VH, Wahlfors T, Saaristo L, Rantapero T, Pelttari LM, Kilpivaara O, Laasanen SL, Kallioniemi A, Nevanlinna H, Aaltonen L, Vessella RL, Auvinen A, et al. HOXB13 G84E mutation in Finland: population-based analysis of prostate, breast, and colorectal cancer risk. Cancer Epidemiol Biomarkers Prev. 2013; 22:452–460. [PubMed: 23292082]
4. Xu J, Dimitrov L, Chang BL, Adams TS, Turner AR, Meyers DA, Eeles RA, Easton DF, Foulkes WD, Simard J, Giles GG, Hopper JL, 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; 77:219–229. [PubMed: 15988677]
5. Lange EM, Robbins CM, Gillanders EM, Zheng SL, Xu J, Wang Y, White KA, Chang BL, Ho LA, Trent JM, Carpten JD, Isaacs WB, 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. [PubMed: 17120048]
6. Pierce BL, Friedrichsen-Karyadi DM, McIntosh L, Deutsch K, Hood L, Ostrander EA, Austin MA, Stanford JL. Genomic scan of 12 hereditary prostate cancer families having an occurrence of pancreas cancer. Prostate. 2007; 67:410–415. [PubMed: 17192958]
7. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, Rafnar T, Gudbjartsson D, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet. 2007; 39:977–983. [PubMed: 17603485]
8. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet. 2008; 40:316–321. [PubMed: 18264097]
9. Cropp CD, Simpson CL, Wahlfors T, Ha N, George A, Jones MS, Harper U, Ponciano-Jackson D, Green TA, Tammela TL, Bailey-Wilson J, Schleutker J. 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; 129:2400–2407. [PubMed: 21207418]
10. 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, 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; 6:4810–4815. [PubMed: 11156239]
11. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, Joensuu H, Raivio T, et al. Comparison of solution-based exome capture

methods for next generation sequencing. Genome Biol. 2011; 12:R94. 2011-12-9-r94. [PubMed: 21955854]

12. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Research. 2010
13. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010; 7:575–576. [PubMed: 20676075]
14. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
15. Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. PON-P: Integrated predictor for pathogenicity of missense variants. Hum Mutat. 2012
16. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011; 39:D945–D950. [PubMed: 20952405]
17. Maqungo M, Kaur M, Kwofie SK, Radovanovic A, Schaefer U, Schmeier S, Oppon E, Christoffels A, Bajic VB. DDPC: Dragon Database of Genes associated with Prostate Cancer. Nucleic Acids Research. 2010
18. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011; 39:D685–D690. [PubMed: 21071392]
19. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27–30. [PubMed: 10592173]
20. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008; 6:e184. [PubMed: 18651794]
21. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, et al. Ensembl 2013. Nucleic Acids Res. 2013; 41:D48–D55. [PubMed: 23203987]
22. 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 Hum Genet. 2007; 81:559–575. [PubMed: 17701901]
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. [PubMed: 22955616]
24. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22:1790–1797. [PubMed: 22955989]
25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
26. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. 2010-11-10-r106. Epub 2010 Oct 27. [PubMed: 20979621]
27. Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, Ghoussaini M, Luccarini C, Dennis J, Jugurnauth-Little S, Dadaev T, Neal DE, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013; 45:385–391. 391e1–391e2. [PubMed: 23535732]
28. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, Eleftheriadis M, Sinning CR, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010; 5:e10693. [PubMed: 20502693]
29. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–1224. [PubMed: 17873874]

30. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N, Nickerson E, Chae SS, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012; 44:685–689. [PubMed: 22610119]

31. Barbieri CE, Bangma CH, Bjartell A, Catto JW, Culig Z, Gronberg H, Luo J, Visakorpi T, Rubin MA. The mutational landscape of prostate cancer. Eur Urol. 2013; 64:567–576. [PubMed: 23759327]

32. Kumar R, Manning J, Spendlove HE, Kremmidiotis G, McKirdy R, Lee J, Millband DN, Cheney KM, Stampfer MR, Dwivedi PP, Morris HA, Callen DF. ZNF652, a novel zinc finger protein, interacts with the putative breast tumor suppressor CBFA2T3 to repress transcription. Mol Cancer Res. 2006; 4:655–665. [PubMed: 16966434]

33. Callen DF, Ricciardelli C, Butler M, Stapleton A, Stahl J, Kench JG, Horsfall DJ, Tilley WD, Schulz R, Nesland JM, Neilsen PM, Kumar R, et al. Co-expression of the androgen receptor and the transcription factor ZNF652 is related to prostate cancer outcome. Oncol Rep. 2010; 23:1045–1052. [PubMed: 20204290]

34. Kumar R, Selth LA, Schulz RB, Tay BS, Neilsen PM, Callen DF. Genome-wide mapping of ZNF652 promoter binding sites in breast cancer cells. J Cell Biochem. 2011; 112:2742–2747. [PubMed: 21678463]

35. Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, Rybicki BA, Isaacs WB, Ingles SA, Stanford JL, Diver WR, Witte JS, et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nat Genet. 2011; 43:570–573. [PubMed: 21602798]

36. Halkidou K, Cook S, Leung HY, Neal DE, Robson CN. Nuclear accumulation of histone deacetylase 4 (HDAC4) coincides with the loss of androgen sensitivity in hormone refractory cancer of the prostate. Eur Urol. 2004; 45:382–389. author reply 389. [PubMed: 15036687]

37. Yang Y, Tse AK, Li P, Ma Q, Xiang S, Nicosia SV, Seto E, Zhang X, Bai W. Inhibition of androgen receptor activity by histone deacetylase 4 through receptor SUMOylation. Oncogene. 2011; 30:2207–2218. [PubMed: 21242980]

38. Chen J, Zhu S, Jiang N, Shang Z, Quan C, Niu Y. HoxB3 promotes prostate cancer cell progression by transactivating CDCA3. Cancer Lett. 2013; 330:217–224. [PubMed: 23219899]

39. Frietze S, Lan X, Jin VX, Farnham PJ. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. J Biol Chem. 2010; 285:1393–1403. [PubMed: 19887448]

40. Huang CN, Huang SP, Pao JB, Chang TY, Lan YH, Lu TL, Lee HZ, Juang SH, Wu PP, Pu YS, Hsieh CJ, Bao BY. Genetic polymorphisms in androgen receptor-binding sites predict survival in prostate cancer patients receiving androgen-deprivation therapy. Ann Oncol. 2012; 23:707–713. [PubMed: 21652578]

41. Biswas AK, Johnson DG. Transcriptional and nontranscriptional functions of E2F1 in response to DNA damage. Cancer Res. 2012; 72:13–17. [PubMed: 22180494]

42. Gulzar ZG, McKenney JK, Brooks JD. Increased expression of NuSAP in recurrent prostate cancer is mediated by E2F1. Oncogene. 2013; 32:70–77. [PubMed: 22349817]

43. Nie Z, Stanley KT, Stauffer S, Jacques KM, Hirsch DS, Takei J, Randazzo PA. AGAP1, an endosome-associated, phosphoinositide-dependent ADP-ribosylation factor GTPase-activating protein that affects actin cytoskeleton. J Biol Chem. 2002; 277:48965–48975. [PubMed: 12388557]

44. Mihara H, Kurihara T, Watanabe T, Yoshimura T, Esaki N. cDNA cloning, purification, and characterization of mouse liver selenocysteine lyase. Candidate for selenium delivery protein in selenoprotein synthesis. J Biol Chem. 2000; 275:6195–6200. [PubMed: 10692412]

45. Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, Hopf C, Huhse B, 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. [PubMed: 14743216]

46. Brandt U. Energy converting NADH:quinone oxidoreductase (complex I). Annu Rev Biochem. 2006; 75:69–92. [PubMed: 16756485]

47. Caldovic L, Morizono H, Gracia Panglao M, Gallegos R, Yu X, Shi D, Malamy MH, Allewell NM, Tuchman M. Cloning and expression of the human N-acetylglutamate synthase gene. Biochem Biophys Res Commun. 2002; 299:581–586. [PubMed: 12459178]

48. Chen JL, Li J, Kiriluk KJ, Rosen AM, Paner GP, Antic T, Lussier YA, Vander Griend DJ. Deregulation of a Hox protein regulatory network spanning prostate cancer initiation and progression. Clin Cancer Res. 2012; 18:4291–4302. [PubMed: 22723371]

49. Kim MS, Je EM, Oh JE, Yoo NJ, Lee SH. Mutational and expressional analyses of SPOP, a candidate tumor suppressor gene, in prostate, gastric and colorectal cancers. APMIS. 2013; 121:626–633. [PubMed: 23216165]

What's new?

The single nucleotide polymorphisms (SNPs) identified by genome-wide association studies explain only a fraction of the familial clustering of prostate cancer (PrCa). In this study, we have exploited next-generation sequencing approaches to uncover less common alleles contributing to PrCa risk. Several novel PrCa-associated variants were identified by targeted re-sequencing of two genomic regions, 2q37 and 17q11.2-q22. RNA sequencing of the selected regions followed by eQTL analysis revealed new relationships between regulatory SNPs and PrCa predisposition.

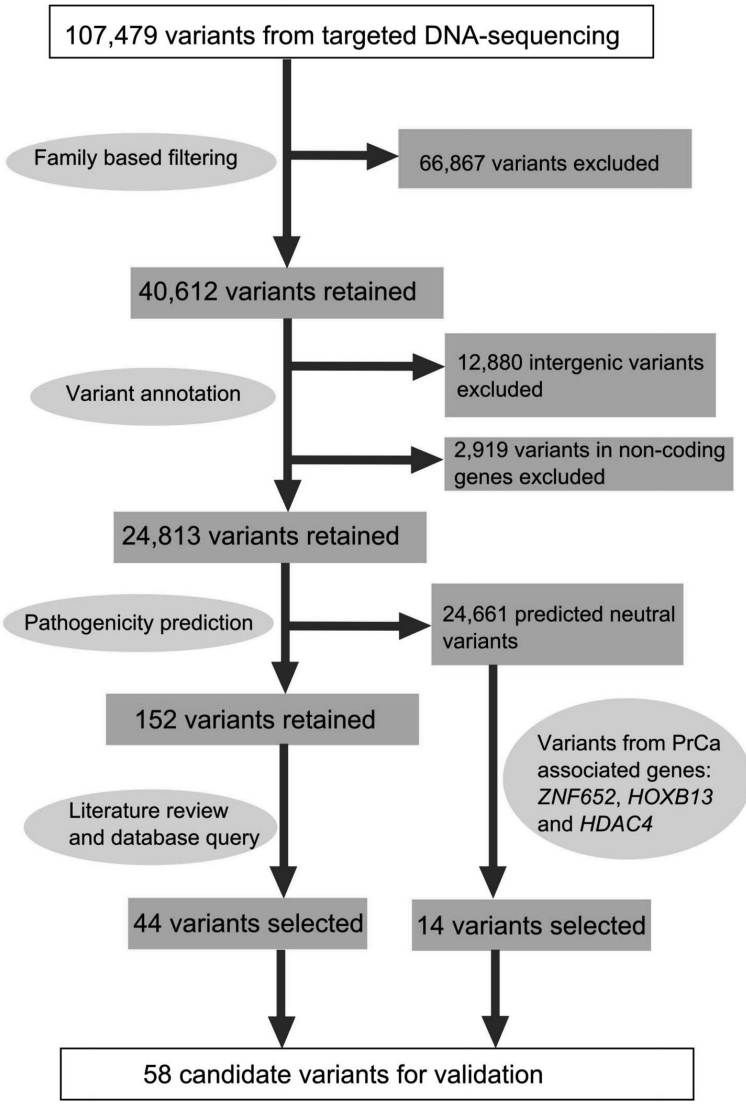

Figure 1. A flowchart describing the variant characterization pipeline

The targeted re-sequencing 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 co-segregate 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 non-pathogenic, were included in the validation because these genes have been associated with PrCa in previous studies.

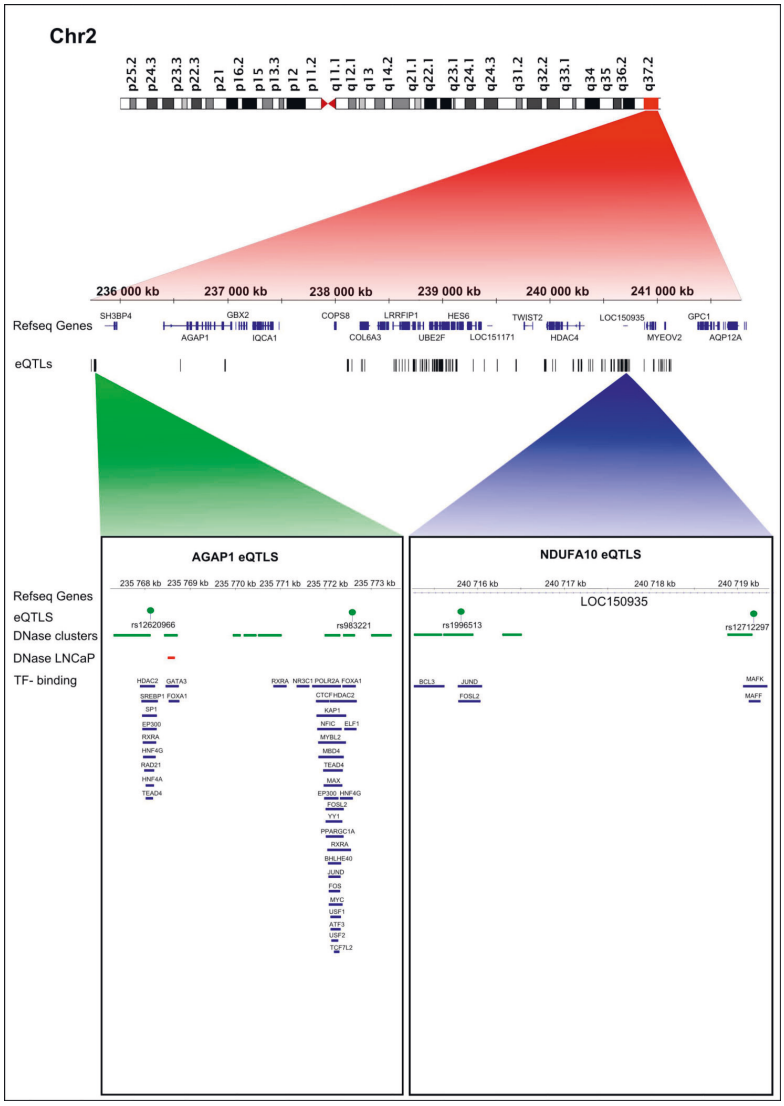

Figure 2. ***Cis*-eQTLs targeting differentially expressed genes on chromosome 2**

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.

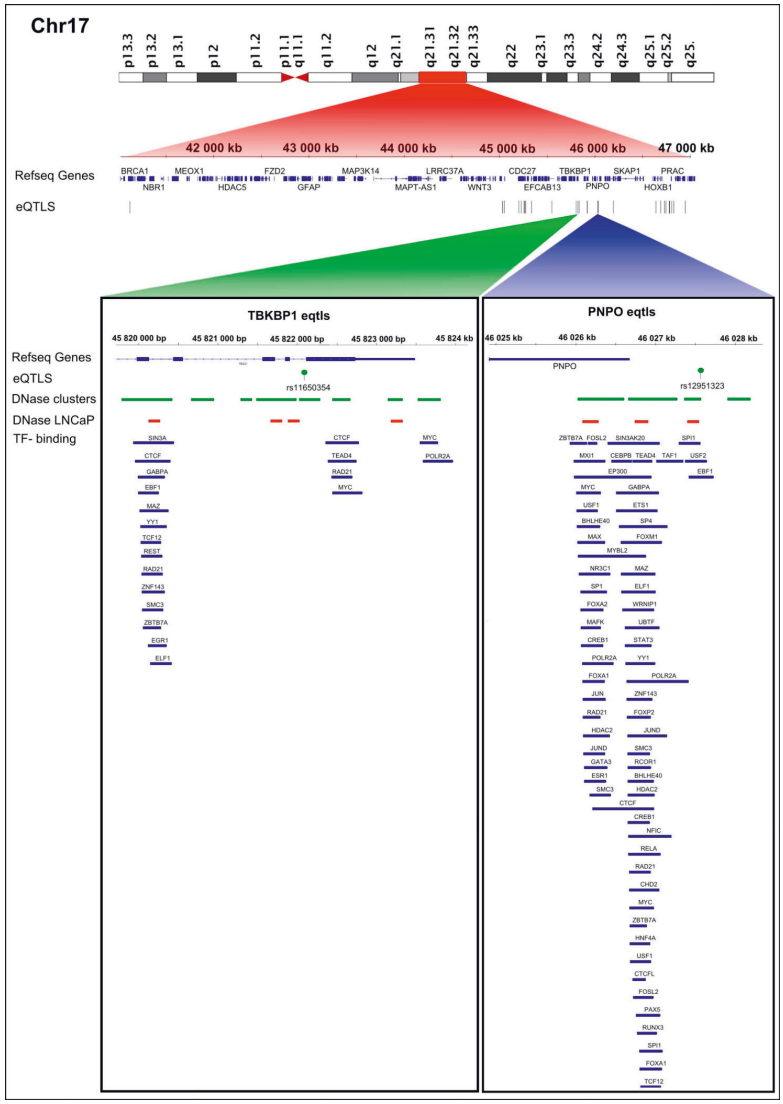

Figure 3. ***Cis*-eQTLs targeting differentially expressed genes on chromosome 17**

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.

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[a] / Maj[b]	F_A[c] / F_U[d] (%)	P value	OR (95% CI)	Pathogenicity prediction[e]
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*	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*	near gene 5'	*ACACA*	17	−A / A	28.8 / 31.1	0.421	0.9 (0.7–1.1)	pathogenic/-/-
rs72828246*	near gene 5'	*ACACA*	17	G / A	28.8 / 30.9	0.459	0.9 (0.7–1.2)	pathogenic/benign/neutral
rs13406410*	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/-/-

[a] Min = minor allele

[b] Maj = major allele

[c] F_A = frequency of the minor allele in cases

[d] F_U = frequency of the minor allele in controls

[e] Pathogenicity prediction results from: Mutation Taster / PolyPhen-2 / Pon-P

* Variants are in linkage disequilibrium.

Chr = chromosome, OR = odds ratio, CI = confidence interval

Bold signifies P < 0.05.

Int J Cancer. Author manuscript; available in PMC 2016 May 15.

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[a] / Maj[b]	F_A[c] / F_U[d] (%)	P value	OR (95% CI)	Pathogenicity prediction[e]
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*	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*	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*	near gene 5'	*MYEOV2*	2	C / A	49.0 / 45.6	**0.037**	1.1 (1.0 – 1.3)	polymorphism/-/-
rs72828246*	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/-/-

[a] Min = minor allele

[b] Maj = major allele

[c] F_A = frequency of the minor allele in cases

[d] F_U = frequency of the minor allele in ontrols

[e] Pathogenicity prediction results from: Mutation Taster / PolyPhen-2 / Pon-P

* Variants are in linkage disequilibrium.

Chr = chromosome, OR = odds ratio, CI = confidence interval

Bold signifies P < 0.05.

Int J Cancer. Author manuscript; available in PMC 2016 May 15.

PUBLICATION II

Whole-exome sequencing of Finnish hereditary breast cancer families

Määttä K, Rantapero T, Lindström A, Nykter M, Kankuri-Tammilehto M, Laasanen SL, Schleutker J.

Eur J Hum Genet. 2016. Jan;25(1):85-93.
doi: 10.1038/ejhg.2016.141.

PUBLICATION III

Integrated RNA-seq and DNase-seq analyses identify phenotype-specific BMP4 signaling in breast cancer

Ampuja M, Rantapero T, Rodriguez-Martinez A, Palmroth M, Alarmo EL, Nykter M, Kallioniemi A.

BMC Genomics. 2017. Jan 11;18(1):68.
doi: 10.1186/s12864-016-3428-1.

Ampuja et al. BMC Genomics (2017) 18:68
DOI 10.1186/s12864-016-3428-1

BMC Genomics

RESEARCH ARTICLE Open Access

Integrated RNA-seq and DNase-seq analyses identify phenotype-specific BMP4 signaling in breast cancer

M. Ampuja[1,2*†], T. Rantapero[1†], A. Rodriguez-Martinez[1,2†], M. Palmroth[1], E. L. Alarmo[1], M. Nykter[1] and A. Kallioniemi[1,2]

Abstract

Background: Bone morphogenetic protein 4 (BMP4) plays an important role in cancer pathogenesis. In breast cancer, it reduces proliferation and increases migration in a cell line-dependent manner. To characterize the transcriptional mediators of these phenotypes, we performed RNA-seq and DNase-seq analyses after BMP4 treatment in MDA-MB-231 and T-47D breast cancer cells that respond to BMP4 with enhanced migration and decreased cell growth, respectively.

Results: The RNA-seq data revealed gene expression changes that were consistent with the in vitro phenotypes of the cell lines, particularly in MDA-MB-231, where migration-related processes were enriched. These results were confirmed when enrichment of BMP4-induced open chromatin regions was analyzed. Interestingly, the chromatin in transcription start sites of differentially expressed genes was already open in unstimulated cells, thus enabling rapid recruitment of transcription factors to the promoters as a response to stimulation. Further analysis and functional validation identified MBD2, CBFB, and HIF1A as downstream regulators of BMP4 signaling. Silencing of these transcription factors revealed that MBD2 was a consistent activator of target genes in both cell lines, CBFB an activator in cells with reduced proliferation phenotype, and HIF1A a repressor in cells with induced migration phenotype.

Conclusions: Integrating RNA-seq and DNase-seq data showed that the phenotypic responses to BMP4 in breast cancer cell lines are reflected in transcriptomic and chromatin levels. We identified and experimentally validated downstream regulators of BMP4 signaling that relate to the different in vitro phenotypes and thus demonstrate that the downstream BMP4 response is regulated in a cell type-specific manner.

Keywords: Bone morphogenetic protein, Breast cancer, NGS, RNA-seq, DNase-seq, Transcription factor

Background

Despite many advances in diagnostics and therapeutics, breast cancer remains the leading cause of cancer death in women [1]. Bone morphogenetic proteins (BMPs) are a group of growth factors that are important players during development [2, 3] but also contribute to cancer formation and progression [4–6]. As a subfamily of the transforming growth factor β (TGF-β) protein superfamily, BMPs are extracellular ligands that bind as dimers to their specific transmembrane receptors and activate the intracellular SMAD signaling pathway leading to phosphorylation of receptor-regulated SMADs (SMAD1/5/9). The activated SMADs bind to SMAD4 and the complex translocates to the nucleus where it regulates the expression of BMP target genes [7, 8]. Alternatively, BMP signals are also mediated through the activation of ERK, JNK and p38 mitogen-activated protein kinase pathways [7, 8].

The functional consequences of BMP signaling depend on the BMP ligand and tissue type. We and others have shown that BMP4 reduces the proliferation of breast cancer cell lines, while simultaneously inducing migration and invasion in a subset of cell lines [9–11]. Similar

* Correspondence: minna.ampuja@uta.fi
†Equal contributors
[1]BioMediTech, University of Tampere, Tampere, Finland
[2]Fimlab Laboratories, Tampere, Finland

dualistic effects upon BMP4 stimulation have also been reported in other tumor types [12]. Concordantly, data from breast cancer patient samples point to a correlation between elevated BMP4 levels and reduced proliferation as well as an increased risk of recurrence [13]. These BMP4-related effects that seem either detrimental (reduced cell growth) or beneficial (increased mobility) for the cancer cells are likely to be mediated by specific BMP4 target genes. The identification of such target genes is thus important since it may allow generation of effective cancer therapies targeting each phenotype independently.

We have previously searched for BMP4 target genes in a set of breast cancer cell lines that predominantly respond to BMP4 treatment by reduction of proliferation [14]. Here, we used next-generation sequencing (NGS) technologies (RNA-seq and DNase-seq) to uncover BMP4-mediated transcriptional events with a specific focus on comparing cells in which BMP4 has opposing effects, namely antiproliferative and promigratory. Out of the nine breast cancer cell lines we have previously studied, T-47D shows one of the most prominent growth reductions and MDA-MB-231 cells display the most overt induction of migration [9, 10], and were thus selected for this study.

RNA-seq method quantifies the level of gene expression across the genome [15] while DNase-seq allows identification of open chromatin regions that are sensitive to digestion by the DNase I endonuclease [16]. Open chromatin regions are considered as sites where transcriptional regulation can take place since they are accessible for regulatory molecules to bind and exert their function. By combining data from RNA-seq and DNase-seq, and using additional data analysis tools, it was possible to identify candidate transcription factors involved in the observed transcriptional responses. This approach thus provides the means to better understand the transcriptional events that link BMP4 signaling and its resulting phenotypes.

Results

We performed RNA-seq and DNase-seq analyses in two breast cancer cell lines, T-47D and MDA-MB-231. The cell lines were treated with BMP4 and vehicle control for 3 h, thus allowing us to specifically focus on early response events. Both vehicle- and BMP4-treated cell lines were sequenced (see methods).

BMP4-elicited transcriptional regulation is highly divergent in the two breast cancer cell lines with different functional responses to BMP4

Sequencing reads from RNA-seq and DNase-seq were aligned to the human genome and further analyzed as described in the methods. To confirm that the two datasets were consistent, we compared the chromatin openness as determined by DNase-seq signal at the transcription start site (TSS) to the expression level of the gene as determined by RNA-seq. As expected, we found that the increased openness of TSS globally correlated with increased gene expression (Additional file 1: Figure S1, Panels A and B). However, the variance is high, indicating that the differences in the chromatin state only partly explain gene expression patterns.

Next we compared the expression levels from RNA-seq between the vehicle- and BMP4-treated cells. This analysis identified 91 differentially expressed genes (DEGs) in MDA-MB-231, of which 58 were upregulated and 33 downregulated (Additional file 2: Table S1). In T-47D, there were 203 DEGs, of which 160 were upregulated and 43 were downregulated (Additional file 3: Table S2). In total, 10 DEGs (*ATOH8*, *BDKRB2*, *BMF*, *GS1-124 K5.4*, *ID1*, *ID2*, *ID3*, *SKIL*, *SMAD6*, and *SMAD9*) were shared by the two cell lines and all of them were upregulated except *GS1-124 K5.4* which was downregulated in both cell lines. To illustrate that BMP4 induces markedly divergent transcriptional responses in these two cell lines, we generated a heatmap to show the expression levels of the protein-coding DEGs (Fig. 1a). Using the DNase-seq data, we examined the chromatin status at the transcription start sites (TSSs) of these protein-coding DEGs. For the majority of the cases the chromatin was open at the TSS before BMP4 stimulation (approximately 86% of all DEGs in both cell lines) (Additional file 1: Figure S1, Panels C and D). For the remaining DEGs, we observed either opening or closing of the TSS after stimulation or no change in the closed chromatin status (Fig. 1a). These data indicate that, at this early time point, the BMP4-induced differential expression mainly involves genes whose transcription does not require changes in the chromatin status at TSS.

The DEG lists included a number of genes involved in the canonical BMP pathway. As expected, *ID1*, *ID2* and *ID3*, known BMP4 target genes, were upregulated in both cell lines (Fig. 1b). Similarly, the receptor-regulated *SMAD9* was upregulated in both cell lines whereas no significant difference in the other receptor-regulated SMADs or *SMAD4* expression was observed. Among the inhibitory SMADs, *SMAD6* was upregulated in both cell lines and *SMAD7* in T-47D. In addition, the BMP type I receptor *BMPR1A* and negative regulators of BMP signaling, *NOG* and *BAMBI*, were upregulated in T-47D while in MDA-MB-231 their expression was not significantly changed (Fig. 1b). Thus BMP4 stimulation leads to expression changes having characteristics of both feedback and feedforward loops.

We then evaluated whether the differentially expressed genes participate in specific biological processes and

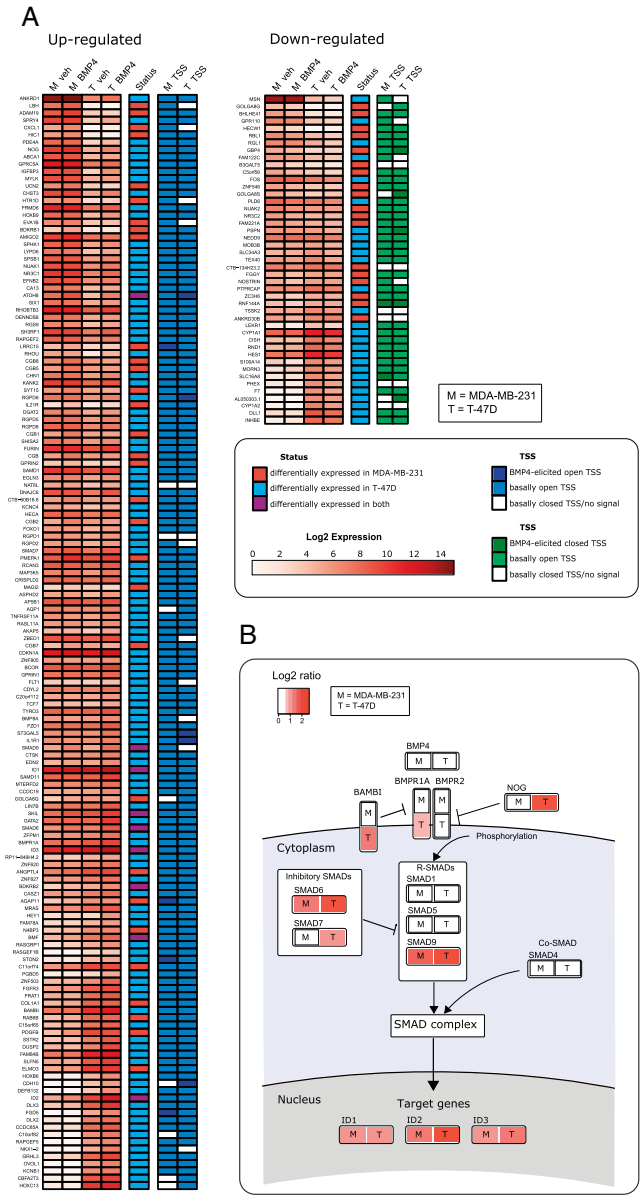

Fig. 1 The RNA-seq and DNAse-seq data reveal cell line-specific responses to BMP4. **a** Gene expression levels of differentially expressed protein-coding genes converted to log2 scale are shown for both cell lines and treatments, upregulated genes on the *left* and downregulated genes on the *right*. The status column denotes the cell line in which the gene is differentially expressed. The rightmost columns indicate the status of the chromatin at transcription start sites (TSS) of the DEGs as measured by DNAse-seq. **b** Illustration of the differentially expressed components of the BMP signaling pathway upon BMP4 treatment

especially assessed whether the non-common DEGs have differing functions. To this end we used DAVID to search for GO terms enriched in the sets of non-common protein-coding DEGs. In MDA-MB-231, most of the enriched terms were related to cell migration whereas organ development and morphogenesis as well as intracellular signaling were the most significant GO terms in T-47D (Table 1). These findings imply that the transcriptional changes are indeed likely to explain the dissimilarities in the phenotypic responses of these two cell lines to BMP4 treatment.

Thereafter, we also wanted to investigate whether the expression levels of DEGs could be linked with survival in breast cancer patients. For this purpose, we used the data publicly available in the TCGA database. The results showed that 20 DEGs in the MDA-MB-231 and 46 DEGs in the T-47D cells associated with either good or poor prognosis (Additional file 4: Tables S3 and S4). Of the nine shared protein-coding DEGs, four (*ATOH8*, *ID3*, *SMAD6* and *SMAD9*) were correlated with survival, all being associated with poor prognosis.

To validate the results of the RNA-seq analysis and to extend the scope of the study beyond the 3 h time point in two cell lines, qRT-PCR was used to study the expression levels of 15 selected DEGs in MDA-MB-231 and T-47D cells as well as in five additional breast cancer cell lines (BT-474, HCC-1954, MCF-7, MDA-MB-361, and MDA-MB-436) and one normal breast epithelial cell line (MCF-10A) treated with BMP4 and vehicle for 3, 6 and 24 h. The genes were selected based on their expression levels and reported cancer association in the literature, and five of these were upregulated according to the RNA-seq in both MDA-MB-231 and T-47D. The expression patterns of the majority of the genes showed similarities across the cell line panel and time points with the clear exception of MDA-MB-436, in which the expression changes were very limited (Fig. 2). Particularly the five shared genes (*ATOH8*, *ID2*, *SKIL*, *SMAD6* and *SMAD9*) as well as *DLX3* were consistently upregulated upon BMP4 treatment throughout the time series thus confirming that they represent common BMP4 target genes. The remaining genes showed more variability with altered expression typically in only two to three cell lines, suggesting that their expression is likely to be influenced by factors that are cell line-specific.

Chromatin landscape and dynamics following BMP4 treatment

To gain more insight into the changes of chromatin structure during BMP4 treatment, we performed peak detection in a genome-wide manner to identify the areas of open chromatin. The peak detection approach was benchmarked by comparison to publicly available DNase-seq data of unstimulated T-47D cell line from ENCODE (see methods), showing that most of the peaks identified in our data are present also in ENCODE samples (Additional file 5: Table S5).

After filtering procedures (see methods), the numbers of identified DNase hypersensitive sites (DHSs) in the MDA-MB-231 cell line were 89,830 and 97,349 in vehicle- and BMP4-treated samples, respectively. In T-47D, the corresponding numbers were 68,000 and 73,881. To obtain a unified set of peaks for both conditions, the overlapping DHSs were merged resulting in a total of 106,154 DHSs in MDA-MB-231 and 110,028 in T-47D. After the merging, the fraction of shared DHSs between BMP4 and vehicle control in MDA-MB-231 samples was 75% while the fraction of unique DHSs in the vehicle was 9% and correspondingly in the BMP4 sample 16% (Additional file 6: Figure S2). In the T-47D cell line,

Table 1 Gene ontology analysis

Cell line	*GO accession*	*GO term*	*Number of genes*	*Adjusted p-value*
MDA-MB-231	GO:0030334	regulation of cell migration	5	2.0×10^{-2}
	GO:0030335	positive regulation of cell migration	4	2.3×10^{-2}
	GO:2000145	regulation of cell motility	5	2.4×10^{-2}
	GO:2000147	positive regulation of cell motility	4	2.5×10^{-2}
	GO:0051272	positive regulation of cellular component movement	4	2.7×10^{-2}
T-47D	GO:0048513	animal organ development	45	2.6×10^{-8}
	GO:0035556	intracellular signal transduction	41	4.5×10^{-8}
	GO:0009887	organ morphogenesis	22	4.0×10^{-7}
	GO:0009966	regulation of signal transduction	36	4.5×10^{-6}
	GO:0007166	cell surface receptor signaling pathway	34	9.5×10^{-5}

The DAVID Functional Annotation Tools was used to reveal significantly enriched GO categories among the differentially expressed protein-coding genes. The analysis was done independently for each cell line and shared differentially expressed genes were omitted. The top five biological function GO terms are shown

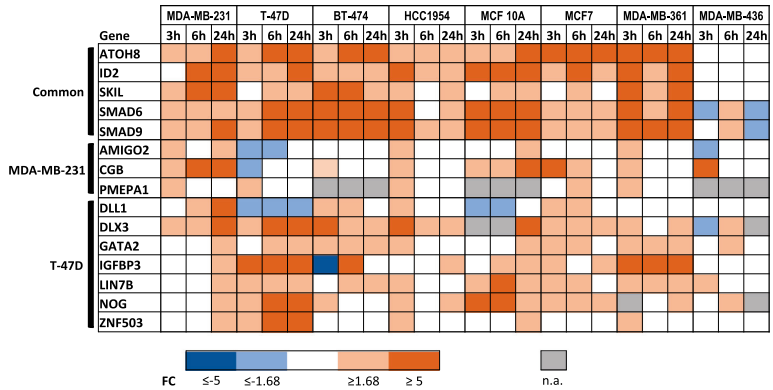

Fig. 2 Expression levels of selected BMP4 target genes by qRT-PCR in a breast cancer cell line panel. The expression levels of 15 DEGs were measured after 3, 6 and 24 h of BMP4 treatment in the indicated cell lines. The color code illustrates the relative expression levels in the BMP4-treated sample as compared to the corresponding vehicle control. FC = Fold change, n.a. = mRNA level too low to allow reliable measurement

the fraction of shared DHSs between the two conditions was 27% whereas the fraction of unique DHSs in the vehicle was 34% and in the BMP4 sample 39% (Additional file 6: Figure S2).

Annotation of the merged DHSs to genomic features revealed a similar distribution in the two cell lines in the vehicle-treated condition, with the largest fraction (>30%) of DHSs locating in introns (Fig. 3a). When comparing the distributions of the BMP4-induced DHSs between the cell lines apparent resemblances were also observed. In both cell lines, the proportion of DHSs associated with intronic and intergenic regions increased after BMP4 stimulation with a corresponding decrease at other genomic locations, including the promoter regions (Fig. 3b).

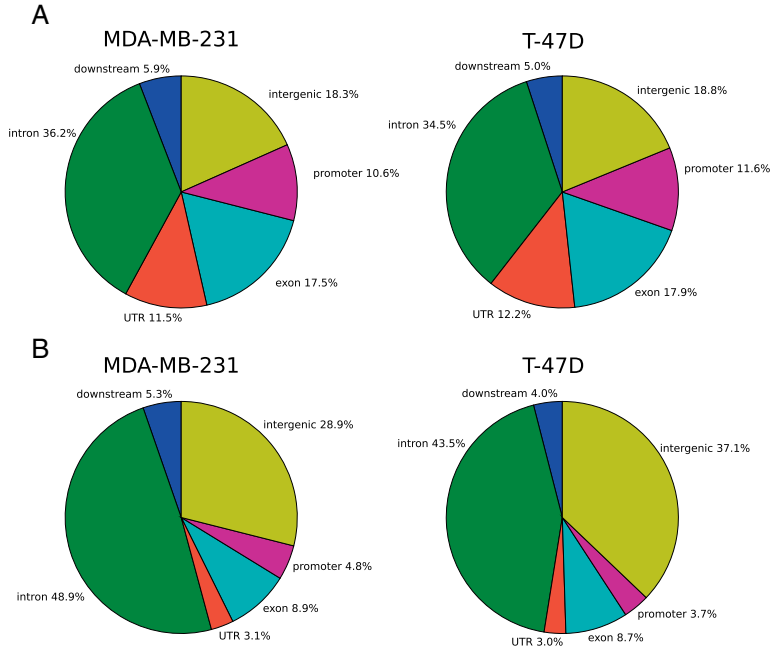

Fig. 3 Distribution of open chromatin regions. Annotation of open chromatin regions in MDA-MB-231 and T-47D after (**a**) vehicle treatment (basal openness) and (**b**) BMP4 treatment (consisting only of the chromatin that opened after BMP4 treatment)

To assess the functional impact of the BMP4-induced global changes in the chromatin structure we conducted an enrichment analysis using GREAT [17] which maps the DHSs to putative regulatory regions of genes and conducts a gene ontology enrichment analysis. The results highlighted e.g. cell motility and organ morphogenesis as enriched biological functions for MDA-MB-231 and T-47D, respectively (Additional file 7: Tables S6 and S7). These results are consistent with those obtained by enrichment analysis of the differentially expressed genes from RNA-seq (Table 1) and thereby suggest that, together with specific target genes, BMP4-induced changes at chromatin level may contribute to the emergence of the different BMP4-mediated phenotypes.

Transcription factor binding site enrichment analysis in open chromatin regions of promoters reveals transcription factors involved in BMP4 signaling regulation

Based on our TSS openness analysis (Fig. 1a), a dominant feature of our data is that the chromatin of the putative BMP4 target genes (identified by RNA-seq) is open already in vehicle-treated cells. This is further supported by our genome-wide peak analysis, where the promoter regions were not overrepresented after the treatment (Fig. 3b). Therefore, the alterations in the chromatin state only partially explain gene expression differences induced by the BMP4 treatment. However, differential transcription factor binding to open promoters may explain the different responses in the cell lines. Therefore we performed transcription factor (TF) motif binding analysis. To assess which TFs might be regulators of the BMP4 response, the sequences of open chromatin sites in the proximal promoters of upregulated genes were analyzed with a total of 426 position weight matrixes (PWMs), representing 401 individual TFs or TF-complexes (see methods). For each TF we calculated an enrichment score (see methods) for the number of binding sites in either MDA-MB-231 or T-47D cells.

This analysis led to the identification of candidate regulator TFs, including multiple members of the SMAD family of TFs, as expected, as well as a number of shared common regulator TFs. To focus on biologically relevant candidates, we filtered out those TFs that were not expressed based on our RNA-seq data. In addition, we included only those TFs whose binding sites (TFBSs) in open chromatin regions of the promoters of DEGs were enriched in one and depleted in the other cell line. The top 15 TFs that are expressed in both cell lines but have a high enrichment score only in one of the cell lines are listed in Tables 2 and 3. Examples of target gene promoters with binding motifs for predicted TFs are shown in Fig. 4a.

For more in-depth functional analysis we selected particular TFs from the top enriched candidates using the following criteria: 1) a binding motif with a quality category of A-C in the HOCOMOCO database, 2) relevance in the context of our model based on literature, 3) not a highly common regulator or part of a large TF family, and 4) high expression level of the TF (>1000 reads) in at least one cell line and differential expression between cell lines according to the RNA-seq. The

Table 2 Top 15 transcription factors enriched in MDA-MB-231 cells

TF name	*Motif*	*Selection by:*	*TF binding sites*	*Ref. sites*	*Expected sites in ref.*	*Ratio of enrichment*	*Mean read count*
MYBL2	MYBB_f1	2, 3, 4	12	2930	6.2	1.92	2197
BACH1	BACH1_si	1, 2, 3	15	3904	8.3	1.81	531
MYC	MYC_f1	1, 2, 4	10	2698	5.7	1.74	3044
MAFK	MAFK_si	2, 3	16	4428	9.4	1.70	688
RELA	TF65_f2	1, 2, 4	19	5467	11.6	1.63	1398
PPARA	PPARA_f1	1, 2, 3	9	2747	5.8	1.54	185
NFIA/B/C/X	[a]	1, 2, 3	15	4669	9.9	1.51	[b]
NFIL3	NFIL3_si	1, 2, 3	11	3494	7.4	1.48	474
FOXA2	FOXA2_f1	1, 2, 3	36	11477	24.4	1.47	434
REL	REL_do	1, 2, 3	17	5422	11.5	1.47	69
ZFHX3	ZFHX3_f1	2, 3	46	14683	31.2	1.47	66
RXRB	RXRB_f1	1, 2, 4	20	6414	13.6	1.47	1015
SMARCC1	SMRC1_f1	1, 4	20	6443	13.7	1.46	1478
ETV5	ETV5_f1	2, 3	16	5199	11.1	1.45	641
NR3C1	GCR_si	1, 2, 4	15	4910	10.4	1.44	1087

The ratio of enrichment is the result of dividing the number of TF binding sites by the number of expected sites. Motifs are derived from the HOCOMOCO database. [a]NFIA + NFIB + NFIC + NFIX_f2, [b]Read count range (51, 148, 748, 444, respectively). *Ref.* reference

Table 3 Top 15 transcription factors enriched in T-47D cells

TF name	*Motif*	*Selection by:*	*TF binding sites*	*Ref. sites*	*Expected sites in ref.*	*Ratio of enrichment*	*Mean read count*
MBD2	MBD2_si	1, 2, 3, 4	101	6664	39.6	2.55	571
TFAP2A	AP2A_f2	1, 2, 3	115	10363	61.6	1.87	941
E4F1	E4F1_f1	2, 3	18	1750	10.4	1.73	310
SP1	SP1_f1	1, 2	392	41453	246.3	1.59	838
CUX1	CUX1_f1	1, 2, 3	13	1462	8.7	1.50	141
E2F2	E2F2_f1	1, 2	17	1941	11.5	1.47	215
AHR	AHR_si	1, 2, 3	9	1030	6.1	1.47	791
SP2	SP2_si	1, 2	140	16512	98.1	1.43	672
CREB1	CREB1_f1	1, 2, 3	23	2720	16.2	1.42	177
CBFB	PEBB_f1	1, 2, 3, 4	46	5461	32.4	1.42	457
ZIC2	ZIC2_f1	1, 2, 3	46	5487	32.6	1.41	118
ZFX	ZFX_f1	1, 2, 3	127	15650	93.0	1.37	287
HIF1A	HIF1A_si	1, 2, 3, 4	15	1890	11.2	1.34	1847
E2F3	E2F3_si	1, 2, 3	16	2019	12.0	1.33	322
XBP1	XBP1_f1	1, 3, 4	12	1545	9.2	1.31	22744

The ratio of enrichment is the result of dividing the number of TF binding sites by the number of expected sites. Motifs are derived from the HOCOMOCO database. *Ref.* reference

last criteria was used to ensure methodological success in subsequent functional assays. With the criteria described above CBFB, HIF1A, and MBD2 were selected for further study. Of these, MBD2 had a large number of binding sites in the promoters of our DEGs while binding sites of the other two TFs were less widespread. In addition, SMAD4 was used as a positive control.

As SMAD4 is a known regulator of BMP signaling, we performed co-occurrence analysis of the binding sites between our three candidate TFs and the SMAD motifs. We found that the MBD2 motif was significantly co-localized with the GC-rich SMAD4 consensus motifs CGCC ($P = 1.1e-9$), GCCGnCGC ($P = 1.3e-14$), and GGCGCC ($P = 2e-10$). As binding sites for CBFB or HIF1A were less frequent across DEGs, statistical significance for co-localization with SMAD motifs could not be reliably evaluated. However, we did find several promoters where SMAD binding sites co-localized with these factors.

Silencing of selected TFs (SMAD4, CBFB, HIF1A, and MBD2) was then used to further evaluate their impact on BMP4 signaling. After 48 h of silencing, the cells were treated with BMP4 for 24 h and the mRNA levels of the validated DEGs were measured to assess whether the silencing influences BMP4 target gene expression (Fig. 4b and Additional file 8: Figure S3). Downregulation of SMAD4 was able to reverse the BMP4-mediated change in the expression of all the tested target genes in both MDA-MB-231 and T-47D cells (Fig. 4c) indicating that these expression changes are indeed transmitted via the canonical BMP pathway. For most of the target genes, MBD2 silencing led to abrogation of the BMP4-mediated induction in gene expression in both cell lines. In T-47D cells, similar data was also obtained for most of the genes upon CBFB (9/10) and HIF1A depletion (6/10). However in MDA-MB-231, silencing of HIF1A resulted exclusively in upregulation of the target genes and both enhanced and diminished expression was seen after CBFB downregulation. Of note, silencing of all of the TFs in T-47D cells led to the enhanced expression of the *DLL1* gene, which was consistent with it being downregulated upon BMP4 treatment. These data imply that the TFs may function as either repressors or enhancers of BMP4 target gene expression in a context-dependent manner.

Discussion

We have previously characterized transcriptional responses of breast cancer cell lines to BMP4 by using microarray technology [14]. However, in that study we focused only on cells that respond to BMP4 by reduced proliferation. Efforts by others to examine BMP signaling target genes have concentrated exclusively on non-cancerous cells [18–20]. Here we set out to uncover the transcriptional responses of breast cancer cell lines with different phenotypes by using one cell line that responds to BMP4 by reduced proliferation (T-47D) and another that reacts with increased migration (MDA-MB-231). Being able to uncover the mechanisms of these two different responses is essential for the understanding of the role of BMP4 in breast cancer pathogenesis. To this end, we used a substantially new approach of combining DNase-seq, RNA-seq and functional experiments.

In order to find the early mediators of BMP4 response, we treated the cells with BMP4 or vehicle control for

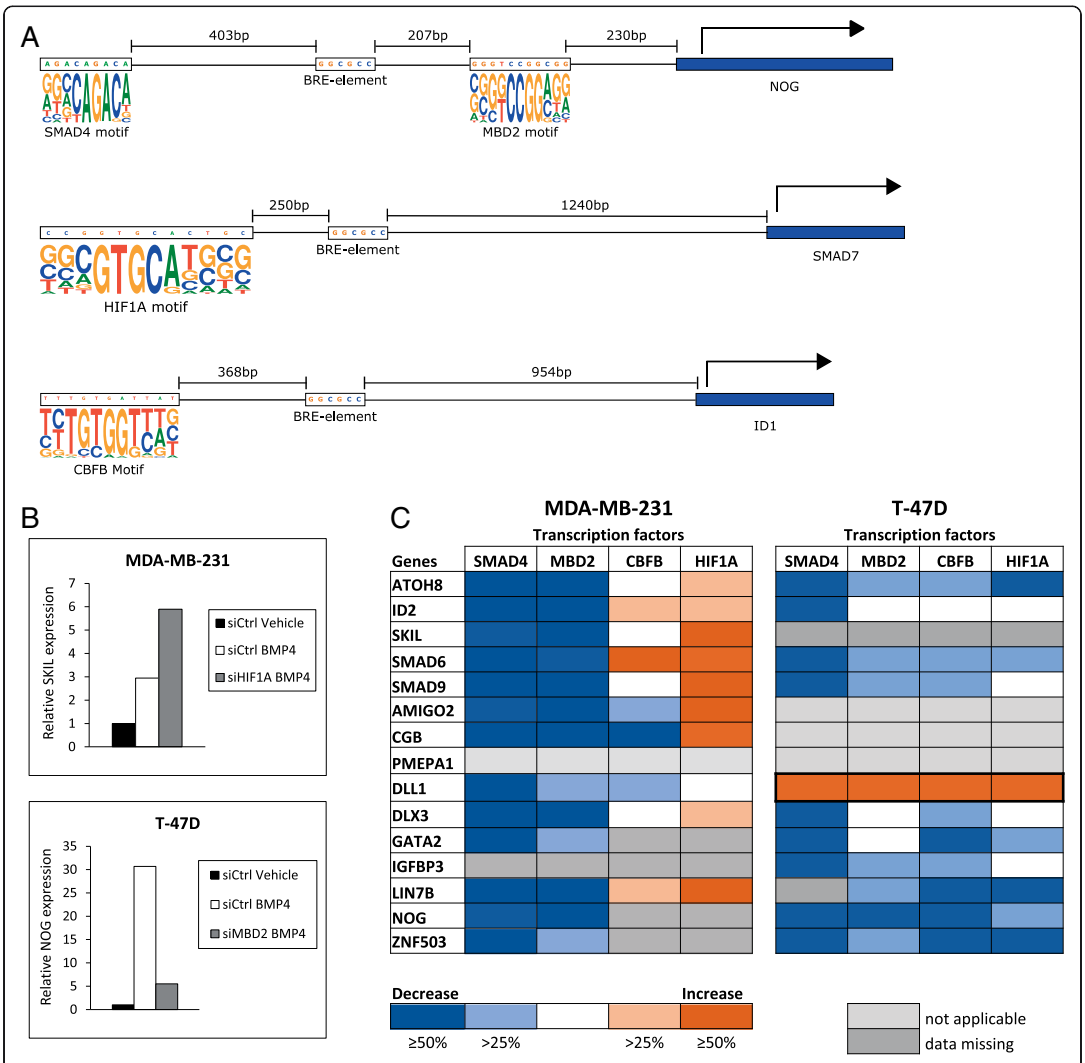

Fig. 4 Examples of predicted TFBSs and the impact of transcription factors on BMP4 target gene expression. **a** The predicted binding sites of transcription factors MBD2, HIF1A and CBFB are depicted at the promoters of *NOG*, *SMAD7* and *ID1* genes, respectively. In addition, known BMP-response elements (BRE) located near the binding sites are illustrated. **b** The TFs were silenced and the cells were treated with BMP4 or vehicle control followed by measurement of target gene expression by qRT-PCR. Examples of relative expression levels of *SKIL* after HIF1A silencing in MDA-MB-231 cells (*top panel*) and *NOG* expression after MBD2 silencing in T-47D cells (*bottom panel*) are shown. **c** Graphical summary of the TF silencing experiments. The order of the genes is identical to that in Fig. 2. Blue color (decreased target gene expression) denotes TFs that were essential for target gene expression and red color those whose silencing led to enhanced target gene expression. Not applicable indicates cases where BMP4 did not alter the baseline gene expression. Data on the DLL1 gene, which is downregulated in T-47D upon BMP4 treatment, are highlighted with a *bold line*

3 h. At this time point, the canonical BMP pathway through SMAD1/5/9 is already activated [9]. The results of RNA-seq revealed that the cell lines responded to BMP4 by upregulating or downregulating a set of genes that were mostly cell line-specific, with only ten common DEGs identified. Consistent with the sequencing data, validation with qRT-PCR across multiple time points (3, 6, and 24 h) and five additional cell lines further confirmed in a wider context the existence of common BMP4 target genes as well as cell line-specific expression patterns. Of the ten shared DEGs, three were known BMP4 target genes (*ID1-3*) and two members of the BMP

signaling pathway (*SMAD6*, *SMAD9*) [3]. The activation of the inhibitory *SMAD6* indicates a negative feedback loop, which in T-47D is reinforced by the upregulation of BMP antagonist *NOG* and the pseudoreceptor *BAMBI*. On the other hand, activation of the receptor-regulated *SMAD9* seems to point to a positive feedback loop, as alongside other R-SMADs, SMAD9 has been found to enhance BMP signaling [21, 22]. However, one study indicated that SMAD9 may have an inhibitory role in BMP signaling [23]. In any case, upregulation of *SMAD9* due to BMP4 treatment has also been recently reported in various cell types, for example in primary fibroblasts, hepatocellular carcinoma and melanoma cells [24].

To understand the function of the cell line-specific DEGs, we used GO analysis to segregate the DEGs into biological process categories and discovered that the results reflected the response of the cell lines to BMP4. Processes related to migration were enriched in the MDA-MB-231 cells, whereas more diverse responses were found in T-47D, including categories comprised of signaling, development and morphogenesis. These findings were corroborated by the DNase-seq data, where we found that BMP4-induced global open chromatin sites were enriched with the same biological categories that were found with RNA-seq data. While categories associated with signaling were observed in both cell lines, in MDA-MB-231 those related to migration were enriched. These data extend our previous results showing enrichment of differentially expressed genes in GO categories that were associated with the BMP4-induced decrease in proliferation [14]. Taken together, the different responses of the cell lines to BMP4 are reflected both at the transcriptional and chromatin levels.

In the analysis of TSS chromatin state we could observe changes in only a few of the genes that were differentially regulated by BMP4. This might be due to the fact that the 3 h stimulation of BMP4 is too short for most of the TSSs to change their chromatin status. Moreover, we could observe that in many cases the chromatin was already open at the TSS, in which case further changes are not needed to enhance the transcriptional activity. Together with the observation that there is a large variation between the chromatin status and gene expression when we extend the analysis to the whole set of protein-coding genes, it can be concluded that the chromatin state of TSSs explains the observed expression patterns only to a small extent. This result was not unexpected, as gene expression is also commonly regulated from regions located far from the TSS, such as enhancers [25, 26].

With genome-wide detection of open chromatin areas we noticed that BMP4 stimulation induces opening of the chromatin mostly in the intronic and intergenic regions. This is consistent with the fact that changes in the TSS and promoter regions were observed with only a few of the differentially expressed genes. Opening of the intronic sequences may indicate increased level of RNA polymerase activity at gene bodies. Chromatin opening at intergenic regions might suggest that additional regulatory control is being attained in large extent through distal regulatory elements such as enhancers and silencers. Thus, already at the early 3-h time point we are able to observe conformational changes that cells may utilize in more detailed regulation of the BMP response. Unfortunately, based on this analysis we were not able to define a specific transcription factor chromatin signature that could be used to define BMP-specific regulatory sequences. Hence detailed analysis of the putative enhancer regions would require more specific measurement data about the chromatin interactions in these cells.

To further characterize the regulation of BMP4 target genes, we analyzed transcription factor binding sites (TFBSs) in the open chromatin regions located on gene promoters. Among the top 15 enriched TFs, there were a few which had previously been linked to BMP target gene regulation. For example, XBP1 and RELA have been shown to be repressors of BMP target genes *Xvent-2* and *Id1*, respectively [27, 28]. Using enrichment of the TFBSs between cell lines as well as other criteria, we selected three TFs (CBFB, HIF1A, and MBD2) for functional characterization and silenced them in the two cell lines. In addition, we used SMAD4, a key component of the canonical BMP pathway, as our positive control and indeed SMAD4 was required for transcriptional regulation of all the BMP4 target genes in the assay. Although BMPs can signal through alternative pathways [7, 8], this result points to regulation through the canonical pathway. In contrast, the response to other transcription factors was more variable and cell line-specific.

MBD2 is a methyl-CpG-binding transcription factor that plays a role in development [29, 30]. Several studies have shown that MBD2 acts as a transcriptional repressor by recruiting co-repressor complexes to promoters, which in turn leads to formation of repressive chromatin through chromatin remodelling [31, 32]. However, there is also evidence that MBD2 can activate transcription by removing methylation from CpG islands located in promoters [33]. In both cell lines, MBD2 seemed to act mainly as an activator of transcription, although its role was more prominent in MDA-MB-231. In our analysis, MBD2 had a large number of binding sites across DEGs and it was highly expressed in both cell lines, consistent with the observed behavior in the silencing experiment. The key role of MBD2 in controlling the BMP4 response suggests that DNA methylation may be involved in BMP4 signaling.

The core-binding factor subunit beta (CBFB) together with the alpha subunit (RUNX1 or RUNX2) is involved in hematopoiesis and skeletal development [34, 35]. It was generally an activator of transcription in T-47D cells where its binding sites were enriched at DEGs promoters, but showed a less constant role in MDA-MB-231. Of note, CBFB has previously been found to influence BMP signaling in chondrocytes [36] and it has also been shown to have invasive properties in breast, prostate and ovarian cancer cells [37, 38].

Hypoxia-inducible factor 1-alpha (HIF1A) is a key regulator of the hypoxia response and has been linked to breast cancer progression [39]. In our binding site enrichment analysis, we observed that HIF1A binding sites are strongly depleted in MDA-MB-231 DEGs although HIF1A has a very high expression in this cell line. We also found that HIF1A was almost exclusively a transcriptional repressor of BMP4 target genes in MDA-MB-231 cells, whereas in T-47D cells it had either no effect or acted as an activator of transcription. Several hypoxia-related genes were found among DEGs, four in MDA-MB-231 cells (*BDKRB2*, *PDGFB*, *ANGPTL-4* and *UCN2*) and three in T-47D cells (*CBFA2T3*, *EGLN3* and *FLT1*). Interestingly, some of these genes have also been linked to cancer progression, for example HIF1A-dependent upregulation of *PDGFB* and *ANGPTL-4* promotes metastasis of hypoxic breast cancer cells [40, 41]. As an additional interesting aspect, HIF1A has been shown to activate BMP4 transcription in pulmonary arterial smooth muscle cells and in murine spleen and ES cells [42–44]. Together these findings support the view that HIF1A is indeed a cell type-specific repressor that controls a particular subset of BMP4-activated target genes.

Conclusions

By combining genome-wide computational analyses and experimental data with functional validation, we were able to extend our knowledge about BMP4 signaling in breast cancer. This study demonstrates that the differential responses to BMP4, reduced proliferation and induced migration, seen in breast cancer cell lines in vitro, are reflected in the expression pattern of BMP4 target genes, thus allowing us to uncover regulatory mechanisms associated with these phenotypes.

By integration of chromatin state and transcription factor binding analyses with gene expression, we were able to identify candidate TFs involved in the regulation of BMP4 response. The function of these TFs was then tested by silencing experiments. From our three candidates, MBD2 emerged as a consistent activator of target gene expression in both cell lines, while HIF1A was shown to act as a repressor in cells with induced migration phenotype and CBFB as an activator, particularly in cells with reduced proliferation phenotype.

While understanding the full complexity of the regulation of BMP4 signaling will require more extensive data, analyses and experiments in wider contexts, our current study established the existence of phenotype-specific BMP response patterns in gene expression. Furthermore, we identified and experimentally validated cell type-specific downstream regulators of BMP signaling that relate to these expression patterns and thus to different in vitro phenotypes.

Methods

Breast cancer cell lines and treatments

Breast cancer cell lines BT-474, HCC-1954, MCF-7, MDA-MB-231, MDA-MB-361, MDA-MB-436, and T-47D as well as the normal immortalized mammary gland cell line MCF-10A were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA) and cultured according to the recommended conditions. The cell lines were authenticated by genotyping and were regularly tested for mycoplasma infection. Cells were seeded, allowed to adhere for 24 h, and treated with 100 ng/ml recombinant human BMP4 protein (R&D Systems, Minneapolis, MN, USA) or vehicle (BMP4 dilution solution). For RNA-seq and DNase-seq, one sample per cell line and treatment was used. Samples were collected 3 h after the treatment, based on our previous results showing SMAD1/5/9 protein phosphorylation [9] and gene expression changes by microarray analyses at this time point [14]. For qRT-PCR, samples representing three biological replicates were collected at indicated time points and pooled.

RNA purification and sequencing library preparation

Total RNA was extracted from BMP4- and vehicle-treated cells using the Absolutely RNA miRNA kit (Agilent Technologies, Palo Alto, CA, USA) according to the manufacturer's instructions. RNA quality was monitored using Agilent 2100 Bioanalyzer (Agilent Technologies). Sequencing libraries were generated using the TruSeq RNA Library Prep kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's directions. Shortly, total RNA was enriched for Poly-A tails and then fragmented. Subsequently, RNA fragments were reverse transcribed into cDNA using random hexamer primers. Then, short fragments were purified and resolved with EB buffer for end reparation and adding poly(A). After that, the short fragments were ligated to sequencing adapters. Finally, suitable fragments were selected for the PCR amplification as templates and separated with agarose gel electrophoresis before sequencing.

Preparation of DNase I-treated DNA and sequencing library

Cells were grown to 80–90% confluency, treated with BMP4 or vehicle for 3 h, and 3×10^7 nuclei were isolated as previously described [16]. DNase I digestion

was performed according to the protocol by Ling and Waxman [45]. First, 7.5×10^6 nuclei were subjected to varying amounts of DNase I and different digestion times in order to optimize the conditions. The qPCR-based DNase hypersensitive site (DHS) cleavage assay [45] was performed using positive control primers surrounding known DHSs in the promoters of housekeeping genes and negative control primers from intergenic insensitive sites (Additional file 9: Table S8). Based on these analyses, 40 units of DNase I for 15 min was selected for the DNase I-treatment. The digestion reaction was followed by phenol-chloroform extraction and size fractionation of DNase I-released fragments by sucrose gradient ultracentrifugation. The DNA fraction with optimal enrichment of DHSs was chosen based on the qPCR-based fragment release assay [45]. Positive control primers were located inside known DHSs in the promoters of housekeeping genes and negative control primers in gene-free regions of different chromosomes (Additional file 9: Table S8). Fraction 7 gave optimal results (DNA fragments less than 1 Kb in size) in all cases and was therefore used for the subsequent steps. Libraries were generated using BGI's in-house protocol. Shortly, 3′-dA overhangs were added and methylated sequencing adaptors were ligated to the DNA fragments. This was followed by PCR amplification and size selection to 200–400 bp, including the adaptor sequence. Undigested DNA from both cell lines was included as an input control.

Deep sequencing

All library construction and deep sequencing steps were carried out at the Beijing Genomics Institute (BGI) (Hong Kong) according to their standard practice. Sequencing was performed on the Illumina HiSeq2000 platform (Illumina). Raw image files were processed by Illumina pipeline for basecalling with default parameters. Reads with too many N bases (>10%) or low base quality (>50% bases with base quality <5) were discarded. On average, we obtained 49 million 90 bp-long paired-end reads from the RNA-seq. For MDA-MB-231 cells, 49,403,872 reads were obtained for the BMP4-treated sample, while 49,424,070 reads resulted from the vehicle-treated sample. The equivalent read amounts for T-47D cells were 49,369,676 and 49,232,294, respectively. Sequencing of the DNase I-digested samples yielded on average 70 million 50 bp-long single-end reads. The specific read amount for MDA-MB-231 cells treated with BMP4 was 70,714,004, and 70,339,810 for vehicle-treated cells. The analogous numbers for T-47D cell line were 79,353,149 and 65,606,222.

Read alignment and normalization of RNA-seq data

RNA-seq reads were aligned using TopHat2 against hg19 reference genome [46]. On average, we were able to align 96% of the reads. For MDA-MB-231 cells, 96.45% of the reads were aligned for the BMP4-treated sample, while 96.55% was the analogous value for vehicle-treated samples. The equivalent numbers for T-47D cell line were 96.33% and 96.02%. Raw expressions were calculated as simple read counts for composite genes constructed from the set of transcripts included in Gencode Genes version 19 [47] using the in-house tool Pypette (https://github.com/annalam/pypette) which is a toolkit built upon Samtools and Bedtools [48, 49]. Read counts were normalized across samples using median of ratios normalization implemented similarly as in DESeq2 R-package [50].

Differential gene expression and GO analysis

In order to find differentially expressed genes (DEG) between BMP4- and vehicle-treated samples, log2 ratios were calculated. Genes having a log2 ratio absolute value of 0.75 or greater were considered differentially expressed. As additional criteria for DEGs, the absolute difference in read counts between the two treatments was required to be at least 50. Functional classification of the differentially expressed protein-coding genes was performed using the DAVID 6.8 version [51, 52].

Survival analysis of DEGs

Each differentially expressed protein - coding gene was tested for association with the survival of breast cancer patients based on the gene expression data obtained from The Cancer Genome Atlas (TCGA) [53, 54]. The patients included in the dataset ($n = 1212$) were divided into low and high expression groups based on the median expression of the gene. The difference in the survival times between the two groups were tested using the log-rank test and Benjamini-Hochberg correction was applied to the P-values. The survival analysis was implemented using the R-package RTCGA toolbox [55].

Read alignment of DNase-seq data and detection of DNase hypersensitive sites (DHSs)

Reads were aligned using bowtie2 [56]. On average, we were able to align 97% of the reads. For MDA-MB-231 cells, 97.49% of the reads were aligned for the BMP4-treated sample, while 97.40% was the analogous value for vehicle-treated samples. The equivalent numbers for T-47D cell line were 97.45% and 97.75%. DNase hypersensitive sites (DHSs) were detected using DFilter [57]. The standard deviation was set to 2, bin size 100 bp and kernel size 50. In addition, the refine parameter was used. To mitigate the effects of mappability and coverage bias samples that had not been treated with DNase I were used as input controls. To remove likely false positives, all DHSs that were covered by less than 20 reads in sample or in input control were omitted from further analysis. Similarly, DHSs located in positions overlapping

blacklisted regions collected by the ENCODE consortium were filtered out [58]. Additionally, adjacent DHSs (distance between peaks 100 bp or less) were merged together. The merged DHSs were annotated by their association to genomic features obtained from Gencode Genes version 19 using Bedtools [49].

Benchmarking the detected DHSs against available ENCODE datasets

To confirm the consistency of the results of our DHS detection in comparison to available ENCODE data, two DNase-seq datasets, each consisting of two T-47D untreated replicates, were retrieved from GEO (accession numbers: GSM816673 and GSM1024762) [59, 60]. The alignment of reads and peak detection were done according to the workflow described above. Further on, we refer to these datasets by their ENCODE biosample identifiers: ENCSR000ELT and ENCSR000EQB.

Finding differential DHSs (ΔDHS) and functional enrichment analysis

In order to describe the change in chromatin hypersensitivity, DHS change scores (ΔDHS) were calculated between the two conditions using a slightly modified formula to the one introduced by He et al. [61]. The DHS change score for i:th DHS was calculated using the following formula:

$$\Delta DHS = \sqrt{n_i^{treated} / \left(\sum_{k=1}^{m} n_k^{treated} \right) / m} - \sqrt{n_i^{vehicle} / \left(\sum_{k=1}^{m} n_k^{vehicle} \right) / m}$$

,where $n_i^{treated}$ is the number of reads mapped to DHS in the treated sample and $n_i^{vehicle}$ is the number of reads mapped to the DHS in the vehicle sample.

The DHSs having ΔDHS equal or greater than 0.20 were selected for enrichment analysis. The analysis was conducted with GREAT version 3.0.0 [17] using the default parameters. The results were ranked and selected based on the binomial test such that all FDR adjusted p-values were required to be less than 0.05. To filter out overly generic ontology terms all categories including more than 1000 genes were filtered out from the final results of the analysis. In addition, too small categories including less than ten genes were removed.

Calculation of DNase coverage of TSS and correlation with gene expression

All possible transcription start sites (TSS), collected from GENCODE transcripts corresponding to protein - coding genes, were extended 1000 bases to both directions. The coverage was calculated for each of these extended TSS regions, which we further refer to simply as TSS. Furthermore, to obtain a single coverage value to describe the openness of the TSS for each protein - coding gene, a weighted sum of the coverages of the TSSs over all the transcripts associated to that gene was calculated. The weight for each transcript's TSS was determined based on the ratio of the estimated expression of the transcript and the total expression of the gene, which was calculated using RSEM [62]. In case the gene was not expressed in either vehicle or stimulated condition, the same ratio which was observed in the other condition was used. Moreover, if the gene was not expressed in either condition, the maximum TSS coverage over all the transcript's TSSs was used as the representative coverage of the TSS of the gene. For visualization purposes, the chromatin status of each gene's TSS was classified into two categories: closed or open. A TSS was considered to be closed if its coverage belonged to the 1. quintile of the TSS coverages of all genes, in that particular cell line and condition. Otherwise the TSS was considered to be open. Each TSS was associated to the corresponding normalized expression value of the gene, which had been obtained by dividing the expression value obtained after median of ratios normalization by the gene's total exon length.

Prediction of transcription factor binding sites in promoters of upregulated genes

In order to find potential transcriptional regulators of BMP4 response, DHSs overlapping proximal promoters (2000 bp upstream regions) of upregulated genes in MDA-MB-231 and T-47D cell lines were scanned with Position Weight Matrices (PWMs). Due to the low signal-to-noise ratio observed in T-47D samples some DHS regions might be narrower or even absent in the data as can be concluded by comparing the promoter-associated DHS regions between our T-47D samples to untreated ENCODE DNase-seq datasets described earlier (see Additional file 5: Table S5). In order to increase the robustness of our analysis we created a composite dataset by taking the union of all promoter-associated DHSs across our samples and all untreated ENCODE samples. The PWMs were created from the curated collection of Weighted Position Count Matrices (WPCMs) retrieved from HOCOMOCO database (version 9) [63]. The PWMs were calculated from weighted matrices of positional counts (WPCM) using the following formula previously introduced by Makeev et al. [64]:

$$S_{b,i} = ln \frac{x_{b,i} + aq_b}{(W + a)q_b}$$

,where $\mathbf{x_{b,i}}$ is the positional count of base **b** in the i:th column of WPCM, **W** is the total weight of the WPCM,

a is the pseudo count defined as **ln(W)** and q_b is the background frequency of base **b** calculated across all the analyzed sequences.

The score for transcription factor binding match **(M_j)** was obtained for each position within the peaks by scanning the sequence using the previously defined PWMs. The score for position **j** when scanning with PWM **S** of length **w** is calculated as follows:

$$M_j = \sum_{i=0}^{w-1} S_{b(i+j),i}$$

We considered a PWM to be a match if the PWM score had a *p*-value less or equal than 0.001. The score thresholds corresponding to the used p-value cut-off were determined using MACRO-APE [65].

Finding enriched and depleted transcription factor binding sites (TFBS) in promoters of upregulated genes

In order to find enriched transcription factor binding sites a background model was generated by selecting the DHSs of all proximal promoters not included in the set of promoters of upregulated genes as the background set. The background set was scanned for transcription factor binding sites as above. Based on the background set, the expected number of transcription factor sites were calculated for the promoter sets of upregulated genes for MDA-MB-231 and T-47D by first dividing the total number of found TFBSs by the cumulative length of the scanned DHSs in the background set and then multiplying this ratio by the cumulative length of the scanned DHSs in the corresponding promoter set of upregulated genes. The ratio of enrichment was then calculated by dividing the observed TFBSs by the number of expected TFBSs.

Co-localization enrichment analysis of selected TFs and known consensus SMAD4-motifs

Six elements including: CAGACA, GTCT, CAGC, CGCC, GGCGCC and GCCGnCGC which have been previously reported as Smad-binding elements (SBEs) [66–69] were selected for co-localization enrichment analysis. The analysis was conducted such that all TFBSs which fall within 200 bp distance of a consensus motif were considered as co-localized with the motif. The binomial test was used to test for enrichment.

qRT-PCR

Quantitative real-time PCR was performed using the Lightcycler 2.0 instrument (Roche, Mannheim, Germany) with LightCycler® TaqMan® Master reaction mix (Roche). Universal probe library (UPL) probes (Roche) and associated primers (Sigma-Aldrich, St. Luis, MO, USA) were used for most of the genes, and the LightCycler FastStart DNA Master SYBR Green I assay (Roche) for the rest. Roche's Reference Gene Assay for HPRT was used for normalization. Primer sequences and probe information are given in Additional file 9: Table S9.

Transcription factor silencing

Transfections to silence the selected TFs in MDA-MB-231 and T-47D cells were performed on 24-well plates using 10 nM siRNA (siGENOME SMARTpool siRNAs, Dharmacon, Lafayette, CO, USA) and either the Interferin reagent (Polyplus-Transfection, SanMarcos, CA, USA) or DharmaFECT (Dharmacon) according to manufacturer's instructions. An ON-TARGETplus Non-targeting Control Pool was used as control (Dharmacon). The knock-down of TFs was confirmed by qRT-PCR and at least 80% reduction in mRNA level was considered as adequate silencing. Forty-eight hours after the transfection, the cells were treated with 100 ng/ml BMP4 or vehicle for 24 h. Cell samples were collected by pooling three identically treated wells and RNA was isolated for subsequent qRT-PCR analyses.

Additional files

Additional file 1: Figure S1. Relationship between chromatin status of TSS and gene expression. The boxplots illustrate the distribution of DNase-seq read coverage at TSS for protein-coding genes at five different levels of gene expression, which were determined by division of expressions into quintiles. Panel A shows the results obtained from untreated MDA-MB-231 cells and panel B the corresponding results for untreated T-47D cells. Panels C and D illustrate the difference between non-expressed and differentially expressed (protein - coding) genes in terms of the chromatin status at TSS in vehicle-treated samples of MDA-MB-231 and T-47D cells, respectively. In both cell lines, chromatin is clearly open at the TSS of differentially expressed genes before the stimulation with BMP4. (PDF 2463 kb)

Additional file 2: Table S1. Differentially expressed genes after BMP4 treatment in MDA-MB-231 cell line. Ensembl IDs, read counts, fold changes and Log2 ratios are shown. The genes are arranged in order from the largest to smallest Log2 ratio, first upregulated genes and then downregulated genes. (XLSX 20 kb)

Additional file 3: Table S2. Differentially expressed genes after BMP4 treatment in T-47D cell line. Ensembl IDs, read counts, fold changes and Log2 ratios are shown. The genes are arranged in order from the largest to smallest Log2 ratio, first upregulated genes and then downregulated genes. (XLSX 34 kb)

Additional file 4: Table S3. Survival analysis of DEGs in MDA-MB-231. **Table S4**. Survival analysis of DEGs in T-47D. Each differentially expressed protein - coding gene was tested for possible association with the survival of breast cancer patients based on the gene expression data obtained from The Cancer Genome Atlas (TCGA). Blue background indicates DEGs shared by both cell lines. Benjamini-Hochberg corrected P-values are shown. No diff. = no association with survival. Not available = not found in TCGA data. (XLSX 16 kb)

Additional file 5: Table S5. Comparison between our T-47D DNase-seq data and analogous data from ENCODE. DNase-seq peaks from promoter regions in BMP4-treated and vehicle-treated T-47D cells were compared to DNase-seq data of T-47D promoters from ENCODE (ENCSR000ELT replicates 1 and 2; ENCSR000EQB replicates 1 and 2). The table shows the percentage of shared peaks between pairs of samples. A high percentage

of the peaks identified in our data are also present in ENCODE samples (cells B6-E7). (XLSX 8 kb)

Additional file 6: Figure S2. Shared DNase-seq peaks between BMP4 and vehicle samples. The number of shared peaks and the number of unique peaks in each treatment group are indicated in the Venn diagram. (PDF 153 kb)

Additional file 7: Table S6. MDA-MB-231 GREAT analysis. **Table S7**. T-47D GREAT analysis. Enrichment of open chromatin peaks using GREAT. Red color denotes categories that are the same as in the RNA-seq ontology analysis. (XLSX 146 kb)

Additional file 8: Figure S3. Transcription factor validation. The chosen TFs were silenced and then treated with BMP4 before measuring target gene expression using qRT-PCR. The transcription factor and cell line in question is stated at the beginning of each page. DLL, which is downregulated in T-47D upon BMP4 treatment, is circled with red. (PDF 256 kb)

Additional file 9: Table S8. Primers used for DNase-seq. **Table S9**. Primer sequences for qRT-PCR based expression analyses of BMP4 target genes and transcription factors. (DOCX 19 kb)

Abbreviations

BMP4: Bone morphogenetic protein 4; DEG: Differentially expressed gene; DHS: DNase hypersensitive site; GO: Gene ontology; PWM: Position weight matrices; TF: Transcription factor; TFBS: Transcription factor binding site; TSS: Transcription start site; WPCM: Weighted position count matrices

Acknowledgements

We thank Kati Rouhento for her skillful technical assistance and Matti Annala for assistance in the RNA-seq data-analysis with Pypette (https://github.com/annalam/pypette).

Funding

Academy of Finland (project no. 269474), the Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (the unit of FimLab), Cancer Society of Finland, Doctoral Programme in Biomedicine and Biotechnology at University of Tampere. MA is supported by the Alfred Kordelin Foundation. ARM is funded by Emil Aaltonen Foundation and Finnish Cultural Foundation.

Availability of data and material

Raw and processed RNA-seq and DNAse-seq data has been deposited to GEO database under accession number GSE84579 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84579).

Authors' contributions

ARM, MN and AK designed the research; ARM, MA, TR, ELA, MP performed research and analyzed data; MA and TR wrote the paper; MN and AK provided critical revisions to the intellectual content and supervised the study. All authors have read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 2 September 2016 Accepted: 16 December 2016
Published online: 11 January 2017

References

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr. Accessed 31 May 2016.
2. Bier E, De Robertis EM. EMBRYO DEVELOPMENT. BMP gradients: A paradigm for morphogen-mediated developmental patterning. Science. 2015;26:348.
3. Brazil DP, Church RH, Surae S, Godson C, Martin F. BMP signalling: agony and antagony in the family. Trends Cell Biol. 2015;25:249–64.
4. Singh A, Morris RJ. The Yin and Yang of bone morphogenetic proteins in cancer. Cytokine Growth Factor Rev. 2010;21:299–313.
5. Alarmo E-L, Kallioniemi A. Bone morphogenetic proteins in breast cancer: dual role in tumourigenesis? Endocr Relat Cancer. 2010;17:R123–39.
6. Ehata S, Yokoyama Y, Takahashi K, Miyazono K. Bi-directional roles of bone morphogenetic proteins in cancer: another molecular Jekyll and Hyde? Pathol Int. 2013;63:287–96.
7. Bragdon B, Beth B, Oleksandra M, Sven S, Daniel K, Joanne J, et al. Bone morphogenetic proteins: a critical review. Cell Signal. 2011;23:609–20.
8. Miyazono K, Kamiya Y, Morikawa M. Bone morphogenetic protein receptors and signal transduction. J Biochem. 2009;147:35–51.
9. Ketolainen JM, Alarmo E-L, Tuominen VJ, Kallioniemi A. Parallel inhibition of cell growth and induction of cell migration and invasion in breast cancer cells by bone morphogenetic protein 4. Breast Cancer Res Treat. 2010;124:377–86.
10. Ampuja M, Jokimäki R, Juuti-Uusitalo K, Rodriguez-Martinez A, Alarmo E-L, Kallioniemi A. BMP4 inhibits the proliferation of breast cancer cells and induces an MMP-dependent migratory phenotype in MDA-MB-231 cells in 3D environment. BMC Cancer. 2013;22:13–429.
11. Guo D, Huang J, Gong J. Bone morphogenetic protein 4 (BMP4) is required for migration and invasion of breast cancer. Mol Cell Biochem. 2012;363:179–90.
12. Kallioniemi A. Bone morphogenetic protein 4-a fascinating regulator of cancer cell behavior. Cancer Genet. 2012;205:267–77.
13. Alarmo E-L, Huhtala H, Korhonen T, Pylkkänen L, Holli K, Kuukasjärvi T, et al. Bone morphogenetic protein 4 expression in multiple normal and tumor tissues reveals its importance beyond development. Mod Pathol. 2013;26:10–21.
14. Rodriguez-Martinez A, Alarmo E-L, Saarinen L, Ketolainen J, Nousiainen K, Hautaniemi S, et al. Analysis of BMP4 and BMP7 signaling in breast cancer cells unveils time-dependent transcription patterns and highlights a common synexpression group of genes. BMC Med Genomics. 2011;25:4–80.
15. Mutz K-O, Kai-Oliver M, Alexandra H, Maren L, Johanna-Gabriela W, Frank S. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol. 2013;24:22–30.
16. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010;2010:pdb.prot5384.
17. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.
18. Fei T, Xia K, Li Z, Zhou B, Zhu S, Chen H, et al. Genome-wide mapping of SMAD target genes reveals the role of BMP signaling in embryonic stem cell fate determination. Genome Res. 2010;20:36–44.
19. Morikawa M, Koinuma D, Tsutsumi S, Vasilaki E, Kanki Y, Heldin CH, et al. ChIP-seq reveals cell type-specific binding patterns of BMP-specific Smads and a novel binding motif. Nucleic Acids Res. 2011;39:8712–27.
20. Genander M, Cook PJ, Ramsköld D, Keyes BE, Mertz AF, Sandberg R, et al. BMP signaling and its pSMAD1/5 target genes differentially regulate hair follicle stem cell lineages. Cell Stem Cell. 2014;15:619–33.
21. Binato R, Alvarez Martinez CE, Pizzatti L, Robert B, Abdelhay E. SMAD 8 binding to mice Msx1 basal promoter is required for transcriptional activation. Biochem J. 2006;393:141–50.
22. Kawai S, Faucheu C, Gallea S, Spinella-Jaegle S, Atfi A, Baron R, et al. Mouse smad8 phosphorylation downstream of BMP receptors ALK-2, ALK-3, and ALK-6 induces its association with Smad4 and transcriptional activity. Biochem Biophys Res Commun. 2000;271:682–7.
23. Tsukamoto S, Mizuta T, Fujimoto M, Ohte S, Osawa K, Miyamoto A, et al. Smad9 is a new type of transcriptional regulator in bone morphogenetic protein signaling. Sci Rep. 2014;4:7596.
24. Katakawa Y, Funaba M, Murakami M. Smad8/9 Is Regulated Through the BMP Pathway. J Cell Biochem. 2016;117:1788–96.
25. Sakabe NJ, Savic D, Nobrega MA. Transcriptional enhancers in development and disease. Genome Biol. 2012;13:238.
26. Symmons O, Spitz F. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. Philos Trans R Soc Lond B Biol Sci. 2013;368:20120358.
27. Hirata-Tsuchiya S, Fukushima H, Katagiri T, Ohte S, Shin M, Nagano K, et al. Inhibition of BMP2-induced bone formation by the p65 subunit of NF-κB via an interaction with Smad4. Mol Endocrinol. 2014;28:1460–70.

28. Cao Y, Knöchel S, Oswald F, Donow C, Zhao H, Knöchel W. XBP1 forms a regulatory loop with BMP-4 and suppresses mesodermal and neural differentiation in Xenopus embryos. Mech Dev. 2006;123:84–96.
29. Klose RJ, Bird AP. Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci. 2006;31:89–97.
30. Menafra R, Stunnenberg HG. MBD2 and MBD3: elusive functions and mechanisms. Front Genet. 2014;5:428.
31. Horike S-I, Cai S, Miyano M, Cheng J-F, Kohwi-Shigematsu T. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet. 2005;37:31–40.
32. Martinowich K, Hattori D, Wu H, Fouse S, He F, Hu Y, et al. DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. Science. 2003;302:890–3.
33. Detich N, Theberge J, Szyf M. Promoter-specific Activation and Demethylation by MBD2/Demethylase. J Biol Chem. 2002;277:35791–4.
34. Okuda T, Takeda K, Fujita Y, Nishimura M, Yagyu S, Yoshida M, et al. Biological characteristics of the leukemia-associated transcriptional factor AML1 disclosed by hematopoietic rescue of AML1-deficient embryonic stem cells by using a knock-in strategy. Mol Cell Biol. 2000;20:319–28.
35. Yoshida CA, Tatsuya F, Takashi F, Ryo F, Naoko K, Shinji K, et al. Core-binding factor β interacts with Runx2 and is required for skeletal development. Nat Genet. 2002;32:633–8.
36. Park N-R, Lim K-E, Han M-S, Che X, Park CY, Kim J-E, et al. Core binding factor β plays a critical role during chondrocyte differentiation. J Cell Physiol. 2016;231:162–71.
37. Davis JN, Rogers D, Adams L, Yong T, Jung JS, Cheng B, et al. Association of core-binding factor β with the malignant phenotype of prostate and ovarian cancer cells. J Cell Physiol. 2010;225:875–87.
38. Mendoza-Villanueva D, Deng W, Lopez-Camacho C, Shore P. The Runx transcriptional co-activator, CBFbeta, is essential for invasion of breast cancer cells. Mol Cancer. 2010;9:171.
39. Wang W, Wei W, Yi-Fu H, Qi-Kai S, Yong W, Xing-Hua H, et al. Hypoxia-inducible factor 1α in breast cancer prognosis. Clin Chim Acta. 2014;428:32–7.
40. Schito L, Rey S, Tafani M, Zhang H, Wong CC, Russo A, et al. Hypoxia-inducible factor 1-dependent expression of platelet-derived growth factor B promotes lymphatic metastasis of hypoxic breast cancer cells. Proc Natl Acad Sci U S A. 2012;109:E2707–16.
41. Zhang H, Wong CC, Wei H, Gilkes DM, Korangath P, Chaturvedi P, et al. HIF-1-dependent expression of angiopoietin-like 4 and L1CAM mediates vascular metastasis of hypoxic breast cancer cells to the lungs. Oncogene. 2012;31:1757–70.
42. Wang J, Fu X, Yang K, Jiang Q, Chen Y, Jia J, et al. Hypoxia inducible factor-1-dependent up-regulation of BMP4 mediates hypoxia-induced increase of TRPC expression in PASMCs. Cardiovasc Res. 2015;107:108–18.
43. Wu D-C, Paulson RF. Hypoxia regulates BMP4 expression in the murine spleen during the recovery from acute anemia. PLoS One. 2010;5:e11303.
44. Pramono A, Zahabi A, Morishima T, Lan D, Welte K, Skokowa J. Thrombopoietin induces hematopoiesis from mouse ES cells via HIF-1alpha-dependent activation of a BMP4 autoregulatory loop. Ann N Y Acad Sci. 2016;1375:38–51.
45. Ling G, Waxman DJ. DNase I digestion of isolated nulcei for genome-wide mapping of DNase hypersensitivity sites in chromatin. Methods Mol Biol. 2013;977:21–33.
46. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
47. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.
48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
50. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
51. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
52. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
53. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163:506–19.
54. Akbani R, Ng PK, Werner HM, Shahmoradgoli M, Zhang F, Ju Z, et al. A pan-cancer proteomic perspective on the cancer genome atlas. Nat Commun. 2014;5:3887.
55. Samur MK. RTCGAToolbox: a new tool for exporting TCGA Firehose data. PLoS One. 2014;9:e106397.
56. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
57. Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nat Biotechnol. 2013;31:615–22.
58. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
59. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
60. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res. 2012;22:1711–22.
61. He HH, Meyer CA, Chen MW, Jordan VC, Brown M, Liu XS. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res. 2012;22:1015–25.
62. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
63. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. 2013;41:D195–202.
64. Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res. 2003;31:6016–26.
65. Vorontsov IE, Kulakovskiy IV, Makeev VJ. Jaccard index based similarity measure to compare transcription factor binding site models. Algorithms Mol Biol. 2013;8:23.
66. Jonk LJ, Itoh S, Heldin CH, ten Dijke P, Kruijer W. Identification and functional characterization of a Smad binding element (SBE) in the JunB promoter that acts as a transforming growth factor-beta, activin, and bone morphogenetic protein-inducible enhancer. J Biol Chem. 1998;273:21145–52.
67. Kim J, Johnson K, Chen HJ, Carroll S, Laughon A. Drosophila Mad binds to DNA and directly mediates activation of vestigial by Decapentaplegic. Nature. 1997;388:304–8.
68. Nakahiro T, Kurooka H, Mori K, Sano K, Yokota Y. Identification of BMP-responsive elements in the mouse Id2 gene. Biochem Biophys Res Commun. 2010;399:416–21.
69. Zawel L, Dai JL, Buckhaults P, Zhou S, Kinzler KW, Vogelstein B, et al. Human Smad3 and Smad4 are sequence-specific transcription activators. Mol Cell. 1998;1:611–7.

PUBLICATION IV

Inherited DNA Repair Gene Mutations in Men with Lethal Prostate Cancer

Rantapero T, Wahlfors T, Kähler A, Hultman C, Lindberg J, Tammela TL, Nykter M, Schleutker J, Wiklund F.

Genes (Basel). 2020. Mar 14;11(3).
doi: 10.3390/genes11030314.

genes

MDPI

Article

Inherited DNA Repair Gene Mutations in Men with Lethal Prostate Cancer

Tommi Rantapero [1], Tiina Wahlfors [1], Anna Kähler [2], Christina Hultman [2], Johan Lindberg [2], Teuvo L. J. Tammela [1], Matti Nykter [1], Johanna Schleutker [3,4] and Fredrik Wiklund [2,*]

1 Faculty of Medicine and Health Technology, Prostate Cancer Research Center, Tampere University, 33100 Tampere, Finland; tommi.rantapero@tuni.fi (T.R.); tiina.wahlfors@veripalvelu.fi (T.W.); teuvo.tammela@pshp.fi (T.L.J.T.); matti.nykter@tuni.fi (M.N.)

2 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden; anna.kahler@ki.se (A.K.); christina.hultman@ki.se (C.H.); johan.lindberg@ki.se (J.L.)

3 Institute of Biomedicine, University of Turku, 20014 Turku, Finland; johanna.schleutker@utu.fi

4 Department of Medical Genetics, Genomics, Laboratory Division, Turku University Hospital, 20521 Turku, Finland

* Correspondence: fredrik.wiklund@ki.se; Tel.: +46-852483979

Received: 19 February 2020; Accepted: 13 March 2020; Published: 14 March 2020

Abstract: Germline variants in DNA repair genes are associated with aggressive prostate cancer (PrCa). The aim of this study was to characterize germline variants in DNA repair genes associated with lethal PrCa in Finnish and Swedish populations. Whole-exome sequencing was performed for 122 lethal and 60 unselected PrCa cases. Among the lethal cases, a total of 16 potentially damaging protein-truncating variants in DNA repair genes were identified in 15 men (12.3%). Mutations were found in six genes with *CHEK2* (4.1%) and *ATM* (3.3%) being most frequently mutated. Overall, the carrier rate of truncating variants in DNA repair genes among men with lethal PrCa significantly exceeded the carrier rate of 0% in 60 unselected PrCa cases ($p = 0.030$), and the prevalence of 1.6% ($p < 0.001$) and 5.4% ($p = 0.040$) in Swedish and Finnish population controls from the Exome Aggregation Consortium. No significant difference in carrier rate of potentially damaging nonsynonymous single nucleotide variants between lethal and unselected PrCa cases was observed ($p = 0.123$). We confirm that DNA repair genes are strongly associated with lethal PrCa in Sweden and Finland and highlight the importance of population-specific assessment of variants contributing to PrCa aggressiveness.

Keywords: prostate cancer; DNA repair genes; lethal cancer

1. Introduction

Prostate cancer (PrCa), the most common male cancer worldwide, has a wide spectrum of clinical behavior that ranges from decades of indolence to rapid metastatic progression and lethality [1]. PrCa is also among the most heritable human cancers, with 57% of the interindividual variation in risk attributed to genetic factors [2]. Genome-wide association studies (GWAS) have thus far confirmed ~170 susceptibility loci that account for over 30% of the familial relative risk [3]. However, the risk variants identified using case-control designs show little or no ability to discriminate between indolent and fatal forms of this disease [4]. Therefore, studies contrasting patients with more and less aggressive disease and those exploring associations with disease progression and prognosis should be more effective at detecting genetic risk factors for aggressive PrCa with prognostic potential.

Inherited and acquired defects in DNA repair genes are a common hallmark of cancer and, to date, numerous inherited DNA repair gene mutations that increase cancer risk has been identified [5]. In particular, mutations in *BRCA1* and *BRCA2* genes, both associated with several DNA repair pathways, confer a strikingly increased risk of breast and ovarian cancer [6]. In addition, it is now recognized that the downregulation of DNA repair response is necessary for tumor progression into a more aggressive phenotype [5]. Accumulating evidence suggests that pathogenic germline variants in known cancer-predisposing genes such as *BRCA2* can increase the risk of developing PrCa, especially the more aggressive form of the disease [7]. Likewise, several other genes that were initially implicated as high-risk genes in cancers other than PrCa, such as *CHEK2* and *BRIP1,* have subsequently been shown to increase the risk of PrCa as well [8–10]. Recent studies have reported a high carrier rate of inherited DNA repair gene mutations among men with metastatic PrCa (11.8%), significantly exceeding the prevalence (4.6%) among men with localized PrCa [11].

In this study, we evaluated germline variants of DNA repair genes in men who died of PrCa. The aim of our study was to identify and investigate the frequency of pathogenic germline variants in men with the lethal form of the disease.

2. Materials and Methods

2.1. Study Subjects

Genomic DNA from a total of 122 lethal PrCa patients was collected from an ongoing collection of Finnish PrCa patients (TAMPERE, $n = 47$) and the Swedish Cancer of Prostate in Sweden (CAPS, $n = 75$) study. To create an extremely aggressive phenotype, the inclusion criterion for lethal PrCa cases was that the patient should have died due to PrCa before the age of 65. All of the Finnish patients were recruited in the Pirkanmaa Hospital District as part of a hereditary PrCa family collection or through collection of sporadic cases treated at the regional hospital [12]. The Swedish CAPS study is a population-based case-control study that enrolled participants between 2001 and 2003 [13]. An additional 70 PrCa patients from the TAMPERE population, not selected for disease aggressiveness or young age at death (hereby denoted unselected cases), with whole-exome sequencing data available were also included to contrast against the lethal cases. Clinical information, such as clinical stage, pathologic grade, nodal or distant metastases, and diagnostic serum levels of PSA and vital status, including cause of death, was obtained through medical records and national cancer registries. All samples were collected with written and signed informed consent. The project was approved by the research ethics committee at Pirkanmaa Hospital District (R03203), the Finnish National Supervisory Authority of Welfare and Health (5569/32/300/05) and by the ethics committees at the Karolinska Institutet (04-449/4 and 06-381/32).

2.2. Sample Preparation, Sequencing and Genotyping

Genomic DNA was extracted from whole blood by standard methods. For the 122 lethal cases, exome capture was performed using Agilent SureSelect Human All Exon 50 M kit (Agilent Technologies, Inc., Santa Clara, CA, USA) according to standard protocol and sequenced at the Science for Life laboratory (Stockholm, Sweden). Of the 70 unselected cases 25 samples were sequenced by BGI Tech Solutions (Hong Kong, China) with exome capture performed by the SureSelect Human All Exon 50 M kit while the remaining 45 unselected cases were sequenced at Mayo Clinic, Rochester, MN, USA with exome capture performed using Agilent SureSelect Human All Exon 50Mb or V4+UTR kits. At each site samples were sequenced using the Illumina Hiseq (Illumina, Inc, San Diego, CA, USA).

2.3. Sample Quality Control and Variant Calling

The reads were aligned against the hg19 genome build retrieved from UCSC using BWA [14]. BEDtools [15] was used to calculate the genome-wide coverage for each sample where samples with less than 30% of bases covered by at least 20 reads were excluded. The PCR duplicates were marked using PICARD [16], and the base score recalibration was performed using GATK [17]. Subsequently, GATK was used to call the variants and genotypes following the GATK best practices protocol for germline exome-sequencing data [18,19]. The candidate false-positive variants were initially filtered using the variant quality score recalibration procedure using the tranche threshold 99.0. Furthermore, variants having an allele fraction of less than 0.3 or a coverage of less than 12 were filtered out. Finally, variants with a readPosRankSum less than or equal to −1.7 were discarded. The variants were annotated using ANNOVAR [20].

2.4. Variant Prioritization

Variants found in 175 DNA repair genes [21–23] were selected for further analysis. To prioritize variants for validation, we utilized a similar approach to that introduced by Mijuskovic and coworkers [7]. The intergenic and common (minor allele frequency > 0.01) variants were filtered out. The remaining rare variants were classified into two categories: potentially damaging and neutral. The potentially damaging variants were further classified into two categories (Tier 1 and Tier 2) based on their impact. The classification was performed utilizing a database of reported associations of variants to clinical phenotypes (ClinVar) provided by ANNOVAR and two tools for pathogenicity prediction, CADD [24] and REVEL [25], of which the latter is specifically designed for discovery of rare deleterious variants. Moreover, the known protein domains from the UniProt [26] database were utilized to assess the pathogenicity of protein truncating variants.

Those variants that are reported as likely benign or benign in ClinVar were classified as neutral. Protein truncating variants (stopgain, frameshift indels or splicing site altering variants) were classified as Tier 1 variants if they had a CADD phred score ≥ 20. Furthermore, the variants were required to be reported to be pathogenic or likely pathogenic by the ClinVar database or alternatively known to affect a protein domain reported in Uniprot (e.g., occurring before or within a protein domain). All nonsynonymous single nucleotide variants (missense variants) reported to be pathogenic or likely pathogenic by ClinVar or had a CADD phred score ≥ 20 and REVEL score ≥ 0.75 were classified as Tier 2 variants. The same prioritization criteria were applied to both case cohorts. The full workflow including details of the sequencing data analysis is illustrated in Figure 1.

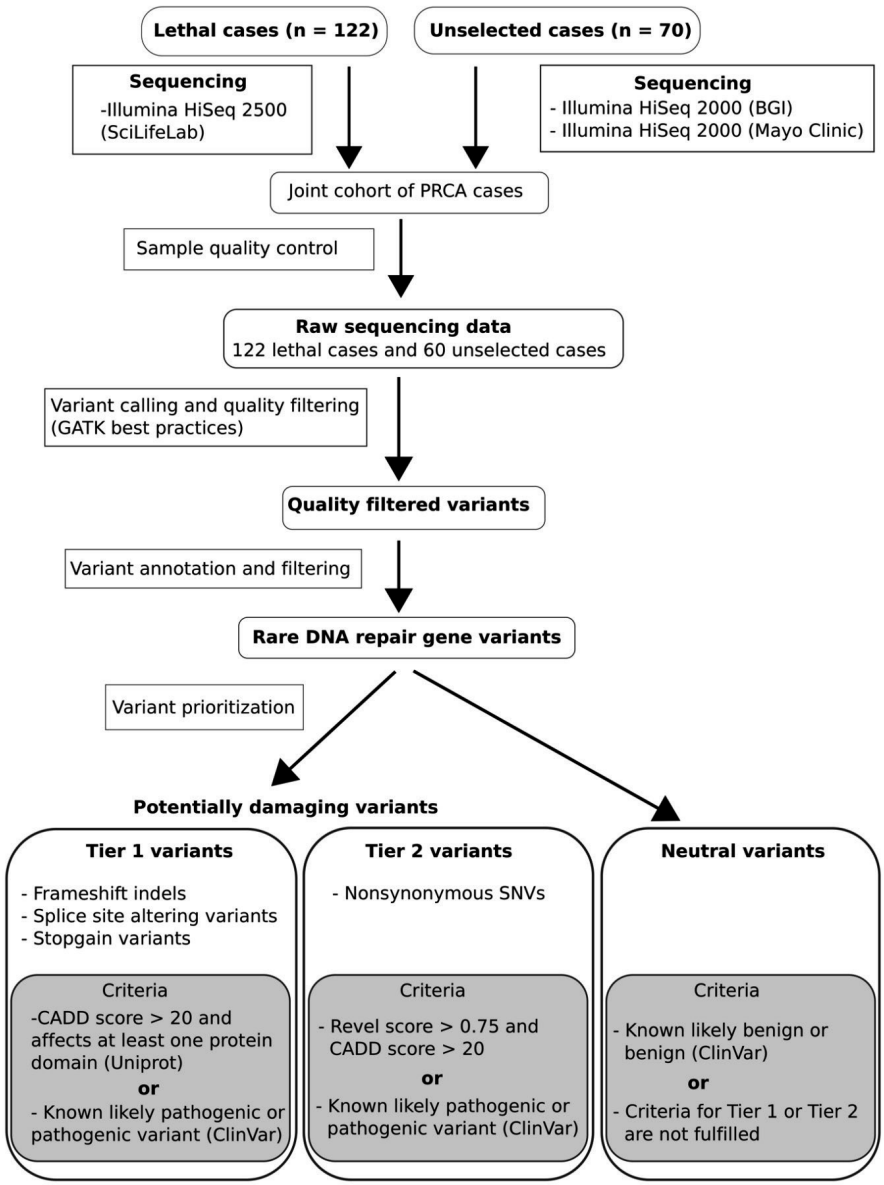

Figure 1. Flow chart describing processing of whole exome sequencing, quality control, variant calling and annotation, and variant prioritizing. PRCA: prostate cancer; ClinVar: database of reported associations of variants to clinical phenotypes; CADD: combined annotation dependent depletion; Revel: rare exome variant ensemble learner.

2.5. Population Frequencies

To explore the expected population allele frequencies of pathogenic variants in the discovered DNA repair genes, we extracted data from two subsets of the Exome Aggregation Consortium (ExAC) browser [27], one set comprising 6192 Swedish population controls and one set comprising 3307 Finnish individuals unselected for cancer history. Full details of the data processing, variant calling

and resources have been described previously [27]. Variant prioritization among these population controls was performed by the same filtering algorithm as described above for the PrCa cases.

2.6. Statistical Analysis

Baseline characteristics were described using the median (interquartile range [IQR]) for continuous variables and absolute and relative frequencies for categorical variables. The frequency of potentially damaging DNA repair gene mutation carriers among the lethal PrCa patients was compared to the frequency in unselected PrCa patients and the two control populations with the use of a two-sided Fisher's exact test. For the control populations, the frequency of mutation carriers in a specific gene was calculated on the basis of the total number of persons for whom sequence coverage was adequate for the given allele, under the assumption that each individual carried at most one deleterious mutation in the explored gene. This assumption may have introduced a slight overestimation in the carrier frequency in the control populations. In all analyses, Tier 1 and Tier 2 mutations were assessed separately. No adjustment was made for multiple testing, and p values less than 0.05 were considered to indicate statistical significance.

3. Results

We performed a comprehensive genetic assessment of DNA repair genes in 122 PrCa cases selected for very aggressive disease and 70 PrCa cases unselected for disease aggressiveness. After exclusion of 10 samples due to insufficient sequencing coverage, 122 lethal cases and 60 unselected cases remained for analysis (Figure 1)—see Table 1 for the clinical characteristics of case cohorts.

Table 1. Clinical characteristics of patients.

	Lethal PrCa (n = 122)	**Unselected PrCa (n = 60)**
Age at diagnosis, median (IQR)	57.0 (55.1–58.2)	66.5 (57.8–73.8)
Diagnostic PSA level (ng/mL), median (IQR)	56.2 (17.9–247.2)	10.8 (7.0–18.8)
Clinical T-stage, n (%)		
TX	2 (1.8)	0 (0.0)
T1	8 (7.3)	20 (38.5)
T2	18 (16.4)	15 (28.8)
T3	61 (55.5)	15 (28.8)
T4	21 (19.1)	2 (3.8)
NA	12	8
Clinical N-stage, n (%)		
NX	86 (78.2)	52 (100.0)
N0	9 (8.2)	0 (0.0)
N1	15 (13.6)	0 (0.0)
NA	12	8
Clinical M-stage, n (%)		
MX	11 (10.0)	14 (26.9)
M0	45 (40.9)	32 (61.5)
M1	54 (49.1)	6 (11.5)
NA	12	8
Gleason score, n (%)		
2–6	11 (10.5)	16 (47.1)
7	36 (34.3)	7 (20.6)
8–10	58 (55.2)	11 (32.4)
NA	17	26
Death due to PrCa, n (%)	122 (100.0)	15 (25.0)
Age at death, median (IQR)	60.0 (57.9–62.9)	79.5 (69.5–84.5)

PrCa: prostate cancer; PSA: prostate-specific antigen; NA: not available.

In total, 22,850,167 variants were discovered and variant prioritization yielded 31 potentially damaging variants distributed across 17 DNA repair genes among the cases (Table 2).

Table 2. Potentially damaging mutations identified in men with lethal prostate cancer.

Gene	RSID	Type	Ref	Alt	Protein Change	ClinVar	CADD/REVEL	MAF	Tier
ATM	rs758081262	stopgain	C	T	Q852X	5	35/-	2.5×10^{-5}	1
ATM	rs761486324	frameshift ins	-	TG	H1082fs	-	-/-	-	1
ATM	rs767099464	frameshift del	C	-	H1083fs	-	-/-	-	1
ATM	rs769142993	missense	G	C	A2524P	4	31/0.89	2.5×10^{-5}	2
ATM	-	frameshift del	AGTAG	-	S2611fs	-	-/-	-	1
ATM	rs753961188	frameshift ins	-	T	L2885fs	5,4	-/-	4.2×10^{-5}	1
ATM	rs376676328	missense	A	G	R2912G	3	29/0.88	3.0×10^{-4}	2
BRCA1	rs41293459	missense	C	T	R1699Q	5,4,3	35/0.79	2.5×10^{-5}	2
CHEK2	rs555607708	frameshift del	G	-	T367fs	5	-/-	1.8×10^{-3}	1
CHEK2	rs137853007	missense	G	A	R145W	5,4	33/0.81	3.3×10^{-5}	2
CHEK2	rs730881700	frameshift ins	-	T	E457fs	5,4	-/-	5.0×10^{-5}	1
CHEK2	rs28909982	missense	T	C	R117G	5,4	27/0.93	1.0×10^{-4}	2
ERCC3	rs753182861	frameshift del	T	-	Q586fs	-	-/-	2.0×10^{-4}	1
ERCC3	rs145267069	missense	A	G	F297S	-	30/0.82	2.5×10^{-5}	2
FAN1	rs778927800	missense	G	A	R749Q	-	34/0.89	8.3×10^{-6}	2
FANCM	rs147021911	stopgain	C	T	Q1701X	4	35/0.12	1.3×10^{-3}	1
HLTF	rs184046773	missense	C	T	G1886A	-	33/0.81	2.0×10^{-4}	2
MRE11A	rs372000848	missense	G	A	R305W	4,3	33/0.85	5.0×10^{-5}	2
MUTYH	rs34126013	missense	G	A	R238W	5,4	33/0.79	9.2×10^{-5}	2
NEIL1	rs5745906	missense	G	A	G169D	-	27/0.86	1.3×10^{-3}	2
NTHL1	rs150766139	stopgain	G	A	Q90X	5,3	35/-	1.5×10^{-3}	1
POLG	rs761584617	missense	G	A	A1115V	-	23/0.80	2.5×10^{-5}	2
POLG	rs113994097	missense	C	G	W748S	5,3	33/0.91	8.0×10^{-4}	2
POLG	rs113994096	missense	G	A	P587L	5,3	28/0.80	1.7×10^{-3}	2
POLG	rs121918052	missense	C	G	Q497H	5,3	26/0.71	2.0×10^{-4}	2
POLL	rs139871590	missense	C	T	G356S	-	34/0.83	1.0×10^{-3}	2
RAD18	rs138830303	stopgain	T	A	K197X	-	36/-	1.0×10^{-4}	1
RECQL	rs149937760	missense	C	T	C414Y	-	33/0.84	2.0×10^{-4}	2
RECQL5	rs768705080	missense	T	G	Y362S	-	32/0.76	8.2×10^{-6}	2
TP53	rs876660754	missense	C	T	V173M	5,4	28/0.89	-	2
TP53	rs779000871	missense	G	A	T170M	3	24/0.87	8.2×10^{-5}	2

Note: ClinVar clinical significance score defines as: 5 = pathogenic, 4 = likely pathogenic, 3 = uncertain significance. Minor allele frequency of variants derived from the Exome Aggregation Consortium. Ref: reference allele; Alt: alternative allele; ClinVar: database of reported associations of variants to clinical phenotypes; CADD: combined annotation dependent depletion; REVEL: rare exome variant ensemble learner; MAF: minor allele frequency; ins: insertion; del: deletion.

Screening of those 17 genes among the population controls revealed 157 potentially damaging variants (Supplementary Table S1) of which 137 were only discovered in the control populations, giving a total of 168 potentially damaging variants. In total, 79 of these variants were known to be pathogenic or likely pathogenic according to ClinVar, while the remaining variants were considered potentially damaging due to their truncating effects on protein domains or by having a REVEL score ≥ 0.75 and a CADD score ≥ 20. Of the 168 potentially damaging variants, 47 were classified as Tier 1 variants and 121 as Tier 2 variants. In total, 21 of the 47 Tier 1 variants were stopgain, 16 were frameshift indels, and 10 were splicing site altering variants.

In exploring the final 168 variants among the 122 lethal cases, 15 men (12.3%) carried at least one potentially damaging Tier 1 germline mutation in a DNA repair gene (one man carried two different Tier 1 mutations in the *ATM* gene), which was significantly higher than that observed in unselected cases (0%, $p = 0.003$, Table 3).

Table 3. Carrier rates of potentially damaging mutations, stratified by Tier 1 and Tier 2 classification, in men with lethal prostate cancer, unselected prostate cancer, and population controls.

	Lethal PrCa (n = 122)	Unselected PrCa (n = 60)	p Value	Finnish Controls (n = 3307)	p Value	Swedish Controls (n = 6192)	p Value
Tier 1							
ERCC3, n (%)	1 (0.82)	0	1.000	0	0.036	3 (0.05)	0.075
RAD18, n (%)	1 (0.82)	0	1.000	0	0.036	0	0.019
ATM, n (%)	4 (3.28)	0	0.304	4 (0.12)	<0.001	10 (0.16)	<0.001
FANCM, n (%)	2 (1.64)	0	1.000	89 (2.69)	0.772	44 (0.71)	0.223
NTHL1, n (%)	2 (1.64)	0	1.000	24 (0.73)	0.236	39 (0.63)	0.187
CHEK2, n (%)	5 (4.10)	0	0.173	60 (1.81)	0.080	5 (0.08)	<0.001
All, *n* (%)	15 (12.30)	0	0.003	177 (5.35)	0.004	101 (1.63)	<0.001
Tier 2							
MUTYH, n (%)	0	1 (1.67)	0.330	34 (1.03)	0.633	75 (1.21)	0.406
ERCC3, n (%)	1 (0.82)	1 (1.67)	0.552	5 (0.15)	0.195	4 (0.06)	0.093
HLTF, n (%)	1 (0.82)	0	1.000	20 (0.60)	0.534	9 (0.15)	0.177
POLL, n (%)	1 (0.82)	0	1.000	15 (0.45)	0.441	28 (0.45)	0.433
MRE11A, n (%)	1 (0.82)	0	1.000	0	0.036	0	0.019
ATM, n (%)	2 (1.64)	0	1.000	13 (0.39)	0.098	28 (0.45)	0.114
RECQL, n (%)	1 (0.82)	0	1.000	0	0.036	13 (0.21)	0.239
FAN1, n (%)	1 (0.82)	0	1.000	2 (0.06)	0.103	16 (0.26)	0.283
NEIL1, n (%)	1 (0.82)	0	1.000	3 (0.09)	0.135	16 (0.26)	0.283
POLG, n (%)	5 (4.10)	0	0.173	197 (5.96)	0.555	190 (3.07)	0.429
TP53, n (%)	2 (1.64)	0	1.000	3 (0.09)	0.012	7 (0.11)	0.012
BRCA1, n (%)	1 (0.82)	0	1.000	2 (0.06)	0.103	5 (0.08)	0.111
RECQL5, n (%)	1 (0.82)	0	1.000	3 (0.09)	0.135	1 (0.02)	0.038
CHEK2, n (%)	1 (0.82)	1 (1.67)	0.552	2 (0.06)	0.103	28 (0.45)	0.433
All, *n* (%)	16 (13.11)	3 (5.00)	0.123	299 (9.04)	0.148	420 (6.78)	0.011

PrCa: prostate cancer. P value: the frequency of potentially damaging DNA repair gene mutation carriers among the lethal PrCa patients was compared to the frequency in unselected PrCa patients and the two control populations with the use of a two-sided Fisher's exact test.

No significant difference in the Tier 1 mutation carrier rate was observed between Swedish (13.3%) and Finnish (10.6%, $p = 0.781$) lethal cases. The two most frequently mutated genes were *CHEK2* (4.1%) and *ATM* (3.3%, Table 3, Figure 2). The observed carrier rate of Tier 1 mutations was significantly higher in the lethal cases compared to the prevalence in the Swedish (1.6%, $p < 0.001$) and the Finnish (5.4%, $p = 0.040$) population controls.

The observed carrier rate of potentially damaging Tier 2 germline mutations was higher in the lethal cases (13.1%) compared to that of the unselected cases (5.0%); however, the difference was not statistically significant ($p = 0.123$, Table 3). Compared to Swedish controls (6.8%, $p = 0.011$), a higher mutation rate was observed among the lethal cases; however, there was no statistically significant difference in the carrier rate of Tier 2 mutations between the lethal cases and the Finnish population controls (9.0%, $p = 0.148$). No significant difference in the Tier 2 mutation carrier rate was observed between Swedish and Finnish lethal cases ($p = 0.102$).

No potentially damaging variants, neither Tier 1 nor Tier 2, were observed in the *BRCA2* gene in any of the PrCa cases. In the population controls, we observed a carrier rate of Tier 1 *BRCA2* mutations of 0.68% and 0.64% in Sweden and Finland, respectively.

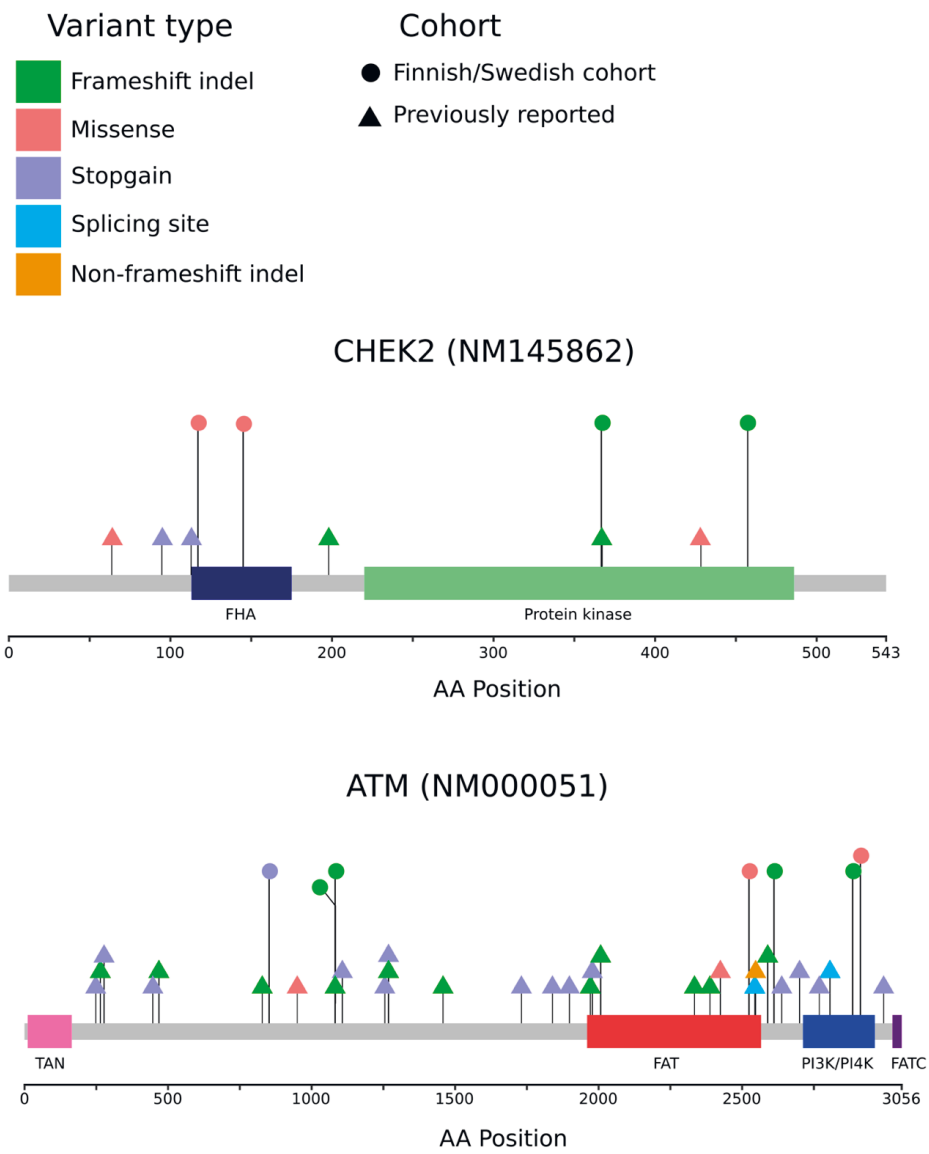

Figure 2. Potentially damaging variants found in the *CHEK2* and *ATM* genes. Locations of variants are shown as lollipop structures. The variants found in the Finnish/Swedish lethal or unselected cases are indicated by circles, and variants found in selected previous studies [7,11,18,28] are indicated by triangles. The variant type is indicated by the color.

4. Discussion

In this study, we characterized the germline variants occurring in the DNA repair pathway from 122 lethal and 60 unselected PrCa patients. In total, 16 potentially damaging protein truncating variants (Tier 1) were identified in 15 men (12.3%) among the lethal cases significantly exceeding the carrier rate of 0% in the unselected cases as well as the population prevalence of 1.6% and 5.4% in Swedish and Finnish population controls. In contrast, the frequency of potentially damaging nonsynonymous

single nucleotide variants (Tier 2) showed similar frequencies among lethal cases, unselected cases and population controls.

Previous studies focusing on aggressive and metastatic PrCa cases have found higher frequencies of deleterious germline variants in *BRCA2* than in any other DNA repair gene and thus considered it to be the major contributor among DNA repair genes to the aggressive phenotype [7,11,29]. However, we observed a frequency of zero pathogenic *BRCA2* variants in our lethal cases, suggesting that *BRCA2* does not play a major role in aggressive and lethal PrCa in the Swedish and Finnish populations. This agrees with earlier studies in which *BRCA1* and *BRCA2* were not found to have a significant contribution to PrCa susceptibility or aggressiveness in Finland or Sweden [30,31]. In a recent study by Mayrhofer and coworkers, sequencing of 217 metastatic PrCa cases from Sweden revealed only two pathogenic *BRCA2* mutation carriers (0.93% carrier rate, [31]). Assuming the same carrier rate among our lethal cases, we would expect to find, on average, 1.1 carriers of *BRCA2* mutations in our study, and our null finding is therefore not surprising. In general, the frequencies of established prostate cancer susceptibility variants deviate from population to population. One such case is the known cancer susceptibility variant G84E in *HOXB13,* which has been shown to have a mutation frequency approximately three-fold higher in Sweden and Finland compared to the mutation frequency in North America [32–34].

ATM and its role in pancreatic cancer was recently reviewed [35] and germline mutations in *ATM* have been associated with predisposition for several cancer forms [36] including PrCa [3]. Several studies have particularly reported potentially damaging variants in *ATM* in aggressive PrCa cases [7,9,29,31]. We also found high frequencies of potentially damaging variants in our lethal cohort (3.28% and 1.64% for Tier 1 and 2 variants, respectively), while in the unselected cases, the frequencies of these variants were found to be very low, similar to those of the population controls. These data support the evidence that deleterious variants in *ATM* are associated with the lethal phenotype of the disease. *ATM* is known to have a predominant role in the DNA damage response, but it also plays a role in maintaining the overall functionality of the cell [37]. *ATM* mutations that cause its inactivation or deficiency have shown a variety of pathological manifestations, including oxidative stress, metabolic syndrome, mitochondrial dysfunction and neurodegeneration. Recently *ATM* deficiency was shown to promote the progression of castration-resistant PrCa by enhancing the Warburg effect, suggesting that *ATM* mutation contributes through a metabolic—in addition to DNA repair—mechanism [38].

CHEK2 variants have been associated with PrCa predisposition in several studies [9,10], and we found that this gene was the most frequently mutated Tier 1 gene in our study (4.1%). In a recent study of 217 metastatic PrCa patients from Sweden [31], *CHEK2* was also the most frequently mutated DNA repair gene (3.8%), highlighting the importance of *CHEK2* mutations for aggressive PrCa in the Nordic population. Of note, in both the present study and the study by Mayrhofer and coworkers [31], c.1100delC was the most commonly observed mutation in *CHEK2* (3.2% and 1.9%, respectively). Wu and coworkers also assessed the frequencies of potentially damaging *CHEK2* variants in lethal cases and in cases with localized low-risk PrCa from the US [39]. Overall, no association was found between *CHEK2* mutation status and lethal disease, but one variant, c.1100delC, was found to have a significantly higher frequency in the lethal cases (1.3%) compared to that of the low-risk PrCa patients (0.2%, $p = 0.004$), supporting the importance of this mutation for lethal PrCa. The c.1100delC has been shown to trigger nonsense-mediated mRNA decay, and subsequent protein analyses suggested that the truncated protein is likely highly unstable [40]. No mechanistic data are available for PrCa, but patients with *CHEK2* mutations are among those showing a high response rate to treatment with the poly-ADP ribose polymerase inhibitor Olaparib when cancers were no longer responding to standard treatments [41].

Of note, only heterozygous carriers of protein-truncating variants were observed in our study conforming to the classical two-hit model for tumor suppressor genes [42,43]. No novel candidate genes within the DNA repair pathway were found in our study. The lack of novel findings is not surprising considering the limited sample size of the study. Moreover, we applied a relatively strict

approach for prioritizing variants, which may have led us to underestimate the role of some genes or even to completely miss potential candidate genes.

We pooled Finnish and Swedish lethal cases to improve the statistical power of the association analysis. No adjustment for possible confounding, for example by population stratification, PSA screening history or family history of PrCa, was performed. Population stratification is always of importance in genetic association studies. However, genotypes from genome-wide single nucleotide polymorphisms were not available for all cases and we were therefore not able to adjust for possible population stratification through principal components in the current study. PSA screening is known to decrease PrCa-specific mortality [44,45] and it is possible that screening history may have confounded our analysis. However, for this to be the case PSA screening history must be associated with carrying pathogenic mutations in DNA repair genes which we find unlikely. Finally, Pritchard and coworkers [11] reported that deleterious mutation frequencies of DNA repair genes did not differ according to whether a family history of PrCa was present among 692 men with metastatic PrCa. Therefore, we argue that confounding by family history is of limited concern in our study.

5. Conclusions

In conclusion, germline variants in DNA repair genes have been shown to be associated with the aggressive form of PrCa—a finding that is supported by our study. Unlike previous studies, we did not observe high numbers of potentially damaging germline variants in *BRCA2*. Instead, mutations in *ATM* and *CHEK2* were found to be most frequent among the lethal cases, highlighting the importance of the population-specific assessment of the variants contributing to the aggressiveness of PrCa.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/3/314/s1, Table S1: Potentially damaging variants discovered in control populations.

Author Contributions: Conceptualization, T.W., J.S. and F.W.; Data curation, T.R. and J.L.; Formal analysis, T.R. and F.W.; Funding acquisition, F.W.; Investigation, J.S. and F.W.; Methodology, T.R., J.L. and M.N.; Project administration, A.K., C.H. and J.L.; Resources, T.W., A.K., C.H., J.L., T.L.J.T., J.S. and F.W.; Software, T.R. and F.W.; Supervision, J.S. and F.W.; Validation, J.S. and F.W.; Visualization, J.S. and F.W.; Writing—original draft, T.R.; Writing—review and editing, T.R., J.S. and F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Sigrid Juselius Foundation (to J.S.), the Academy of Finland (#251074 and #310115 to J.S.), the Cancer Foundation Finland (to J.S. and T.R.), the National Cancer Institute Grant U01 CA 89600 (support for the ICPCG), the Tampere Graduate Program in Biomedicine and Biotechnology (to T.R.), the Swedish Cancer Society (CAN 2016/818), and the Nordic Cancer Union.

Acknowledgments: Kirsi Rouhento is thanked for her assistance.

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

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **2019**, *69*, 7–34. [CrossRef] [PubMed]
2. Mucci, L.A.; Hjelmborg, J.B.; Harris, J.R.; Czene, K.; Havelick, D.J.; Scheike, T.; Graff, R.E.; Holst, K.; Moller, S.; Unger, R.H.; et al. Familial risk and heritability of cancer among twins in nordic countries. *JAMA* **2016**, *315*, 68–76. [CrossRef] [PubMed]
3. Schumacher, F.R.; Al Olama, A.A.; Berndt, S.I.; Benlloch, S.; Ahmed, M.; Saunders, E.J.; Dadaev, T.; Leongamornlert, D.; Anokian, E.; Cieza-Borrella, C.; et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **2018**, *50*, 928–936. [CrossRef] [PubMed]
4. Szulkin, R.; Karlsson, R.; Whitington, T.; Aly, M.; Gronberg, H.; Eeles, R.A.; Easton, D.F.; Kote-Jarai, Z.; Al Olama, A.A.; Benlloch, S.; et al. Genome-wide association study of prostate cancer-specific survival. *Cancer Epidemiol. Biomark. Prev.* **2015**, *24*, 1796–1800. [CrossRef]
5. Jeggo, P.A.; Pearl, L.H.; Carr, A.M. DNA repair, genome stability and cancer: A historical perspective. *Nat. Rev. Cancer* **2016**, *16*, 35–42. [CrossRef]

6. Friedenson, B. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC Cancer* **2007**, *7*, 152. [CrossRef]
7. Mijuskovic, M.; Saunders, E.J.; Leongamornlert, D.A.; Wakerell, S.; Whitmore, I.; Dadaev, T.; Cieza-Borrella, C.; Govindasami, K.; Brook, M.N.; Haiman, C.A.; et al. Rare germline variants in DNA repair genes and the angiogenesis pathway predispose prostate cancer patients to develop metastatic disease. *Br. J. Cancer* **2018**, *119*, 96–104. [CrossRef]
8. Kote-Jarai, Z.; Jugurnauth, S.; Mulholland, S.; Leongamornlert, D.A.; Guy, M.; Edwards, S.; Tymrakiewitcz, M.; O'Brien, L.; Hall, A.; Wilkinson, R.; et al. A recurrent truncating germline mutation in the BRIP1/FANCJ gene and susceptibility to prostate cancer. *Br. J. Cancer* **2009**, *100*, 426–430. [CrossRef]
9. Paulo, P.; Maia, S.; Pinto, C.; Pinto, P.; Monteiro, A.; Peixoto, A.; Teixeira, M.R. Targeted next generation sequencing identifies functionally deleterious germline mutations in novel genes in early-onset/familial prostate cancer. *PLoS Genet.* **2018**, *14*, e1007355. [CrossRef]
10. Seppala, E.H.; Ikonen, T.; Mononen, N.; Autio, V.; Rokman, A.; Matikainen, M.P.; Tammela, T.L.; Schleutker, J. CHEK2 variants associate with hereditary prostate cancer. *Br. J. Cancer* **2003**, *89*, 1966–1970. [CrossRef]
11. Pritchard, C.C.; Mateo, J.; Walsh, M.F.; De Sarkar, N.; Abida, W.; Beltran, H.; Garofalo, A.; Gulati, R.; Carreira, S.; Eeles, R.; et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *N. Engl. J. Med.* **2016**, *375*, 443–453. [CrossRef] [PubMed]
12. Schleutker, J.; Matikainen, M.; Smith, J.; Koivisto, P.; Baffoe-Bonnie, A.; Kainu, T.; Gillanders, E.; Sankila, R.; Pukkala, E.; Carpten, 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*, 4810–4815. [PubMed]
13. Lindmark, F.; Zheng, S.L.; Wiklund, F.; Bensen, J.; Balter, K.A.; Chang, B.; Hedelin, M.; Clark, J.; Stattin, P.; Meyers, D.A.; et al. H6D polymorphism in macrophage-inhibitory cytokine-1 gene associated with prostate cancer. *J. Natl. Cancer Inst.* **2004**, *96*, 1248–1254. [CrossRef] [PubMed]
14. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
15. Quinlan, A.R. BEDTools: The swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinform.* **2014**, *47*, 11–12. [CrossRef] [PubMed]
16. PICARD. Available online: http://broadinstitute.github.io/picard/ (accessed on 6 February 2019).
17. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]
18. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef]
19. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
20. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef]
21. Lange, S.S.; Takata, K.; Wood, R.D. DNA polymerases and cancer. *Nat. Rev. Cancer* **2011**, *11*, 96–110. [CrossRef]
22. Wood, R.D.; Mitchell, M.; Lindahl, T. Human DNA repair genes, 2005. *Mutat. Res.* **2005**, *577*, 275–283. [CrossRef] [PubMed]
23. Wood, R.D.; Mitchell, M.; Sgouros, J.; Lindahl, T. Human DNA repair genes. *Science* **2001**, *291*, 1284–1289. [CrossRef] [PubMed]
24. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [CrossRef] [PubMed]
25. Ioannidis, N.M.; Rothstein, J.H.; Pejaver, V.; Middha, S.; McDonnell, S.K.; Baheti, S.; Musolf, A.; Li, Q.; Holzinger, E.; Karyadi, D.; et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **2016**, *99*, 877–885. [CrossRef] [PubMed]

26. UniProt, C. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* **2019**, *47*, D506–D515. [CrossRef]
27. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B.; et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285–291. [CrossRef]
28. Hart, S.N.; Ellingson, M.S.; Schahl, K.; Vedell, P.T.; Carlson, R.E.; Sinnwell, J.P.; Barman, P.; Sicotte, H.; Eckel-Passow, J.E.; Wang, L.; et al. Determining the frequency of pathogenic germline variants from exome sequencing in patients with castrate-resistant prostate cancer. *BMJ Open* **2016**, *6*, e010332. [CrossRef]
29. Na, R.; Zheng, S.L.; Han, M.; Yu, H.; Jiang, D.; Shah, S.; Ewing, C.M.; Zhang, L.; Novakovic, K.; Petkewicz, J.; et al. Germline Mutations in ATM and BRCA1/2 Distinguish Risk for Lethal and Indolent Prostate Cancer and are Associated with Early Age at Death. *Eur. Urol.* **2017**, *71*, 740–747. [CrossRef]
30. Ikonen, T.; Matikainen, M.P.; Syrjakoski, K.; Mononen, N.; Koivisto, P.A.; Rokman, A.; Seppala, E.H.; Kallioniemi, O.P.; Tammela, T.L.; Schleutker, J. BRCA1 and BRCA2 mutations have no major role in predisposition to prostate cancer in Finland. *J. Med. Genet.* **2003**, *40*, e98. [CrossRef]
31. Mayrhofer, M.; De Laere, B.; Whitington, T.; Van Oyen, P.; Ghysel, C.; Ampe, J.; Ost, P.; Demey, W.; Hoekx, L.; Schrijvers, D.; et al. Cell-free DNA profiling of metastatic prostate cancer reveals microsatellite instability, structural rearrangements and clonal hematopoiesis. *Genome Med.* **2018**, *10*, 85. [CrossRef]
32. Ewing, C.M.; Ray, A.M.; Lange, E.M.; Zuhlke, K.A.; Robbins, C.M.; Tembe, W.D.; Wiley, K.E.; Isaacs, S.D.; Johng, D.; Wang, Y.; et al. Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.* **2012**, *366*, 141–149. [CrossRef] [PubMed]
33. Laitinen, V.H.; Wahlfors, T.; Saaristo, L.; Rantapero, T.; Pelttari, L.M.; Kilpivaara, O.; Laasanen, S.L.; Kallioniemi, A.; Nevanlinna, H.; Aaltonen, L.; et al. HOXB13 G84E mutation in Finland: Population-based analysis of prostate, breast, and colorectal cancer risk. *Cancer Epidemiol. Biomark. Prev.* **2013**, *22*, 452–460. [CrossRef] [PubMed]
34. Xu, J.; Lange, E.M.; Lu, L.; Zheng, S.L.; Wang, Z.; Thibodeau, S.N.; Cannon-Albright, L.A.; Teerlink, C.C.; Camp, N.J.; Johnson, A.M.; 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. [CrossRef] [PubMed]
35. Nanda, N.; Roberts, N.J. ATM serine/threonine kinase and its role in pancreatic risk. *Genes* **2020**, *11*, 108. [CrossRef] [PubMed]
36. Choi, M.; Kipps, T.; Kurzrock, R. ATM mutations in cancer: Therapeutic implications. *Mol. Cancer Ther.* **2016**, *15*, 1781–1791. [CrossRef] [PubMed]
37. Guleria, A.; Chandna, S. ATM kinase: Much more than a DNA damage responsive protein. *DNA Repair (Amst)* **2016**, *39*, 1–20. [CrossRef]
38. Xu, L.; Ma, E.; Zeng, T.; Zhao, R.; Tao, Y.; Chen, X.; Groth, J.; Liang, C.; Hu, H.; Huang, J. ATM deficiency promotes progression of CRPC by enhancing Warburg effect. *Endocr. Relat. Cancer* **2019**, *26*, 59–71. [CrossRef]
39. Wu, Y.; Yu, H.; Zheng, S.L.; Na, R.; Mamawala, M.; Landis, T.; Wiley, K.; Petkewicz, J.; Shah, S.; Shi, Z.; et al. A comprehensive evaluation of CHEK2 germline mutations in men with prostate cancer. *Prostate* **2018**, *78*, 607–615. [CrossRef]
40. Anczukow, O.; Ware, M.D.; Buisson, M.; Zetoune, A.B.; Stoppa-Lyonnet, D.; Sinilnikova, O.M.; Mazoyer, S. Does the nonsense-mediated mRNA decay mechanism prevent the synthesis of truncated BRCA1, CHK2, and p53 proteins? *Hum. Mutat.* **2008**, *29*, 65–73. [CrossRef]
41. Mateo, J.; Carreira, S.; Sandhu, S.; Miranda, S.; Mossop, H.; Perez-Lopez, R.; Nava Rodrigues, D.; Robinson, D.; Omlin, A.; Tunariu, N.; et al. DNA-Repair defects and olaparib in metastatic prostate cancer. *N. Engl. J. Med.* **2015**, *373*, 1697–1708. [CrossRef]
42. Knudson, A.G., Jr. Mutation and cancer: Statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 820–823. [CrossRef] [PubMed]
43. Wang, L.H.; Wu, C.F.; Rajasekaran, N.; Shin, Y.K. Loss of tumor suppressor gene function in human cancer: An overview. *Cell Physiol. Biochem.* **2018**, *51*, 2647–2693. [CrossRef] [PubMed]

44. Schroder, F.H.; Hugosson, J.; Roobol, M.J.; Tammela, T.L.; Ciatto, S.; Nelen, V.; Kwiatkowski, M.; Lujan, M.; Lilja, H.; Zappa, M.; et al. Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.* **2009**, *360*, 1320–1328. [CrossRef] [PubMed]
45. Schroder, F.H.; Hugosson, J.; Roobol, M.J.; Tammela, T.L.; Zappa, M.; Nelen, V.; Kwiatkowski, M.; Lujan, M.; Maattanen, L.; Lilja, H.; et al. Screening and prostate cancer mortality: Results of the European randomised study of screening for prostate cancer (ERSPC) at 13 years of follow-up. *Lancet* **2014**, *384*, 2027–2035. [CrossRef]