Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
GENETIC AND GENOMIC TECHNOLOGIES AND DIAGNOSIS OF AGGRESSIVE PROSTATE CANCER

A Case-Control Study

Venkatesh Vaidyanathan

A thesis submitted in complete fulfilment of the requirements for the degree of Doctor of Philosophy in Health Sciences, The University of Auckland, 2017.
Abstract

Prostate cancer (PCa) is the second-most commonly diagnosed cancer among men worldwide. In the year 2012 itself, approximately 1.1 million men were diagnosed with PCa, accounting for 15% of all the new cancer cases registered in men worldwide. With the alarming estimate that at least 1 in 6 PCa patients is in the risk of developing aggressive form of the disease, the identification of a predictive biomarker for PCa is of much importance. More so for New Zealand’s perspective, because the highest recorded rate of men with PCa relative to the overall population of men is observed in this region in addition to higher mortality rates compared to records from United States of America (USA). Meanwhile, there are a number of unique factors about the environment and lifestyle of the New Zealand population that need to be considered when analyzing various diseases. There are a high number of tobacco smokers in New Zealand, deficiency of trace elements such as selenium in the New Zealand soil impacting selenium nutrition from locally produced food, and New Zealand recording an overweight population. This motivated us to look into various gene x environment interactions and the risk of aggressive PCa.

For this SNP-based study, the PubMed database was screened for research articles based on Genome-Wide Association Studies (GWAS) and/or case-control studies published on or after the year 2000. Attention was given to relevant research papers describing SNP association with PCa among patients with European ethnicity only.

We herewith present the data and analysis of gene x environment interaction and risk of aggressive PCa with 136 single nucleotide polymorphisms (SNPs) present in various genes and undefined regions collected from 197 men with clinically diagnosed aggressive PCa, 57 men clinically diagnosed with non-aggressive PCa and 369 healthy controls. We have identified several SNPs that have risk associations, both, with and without environmental
interactions. These include certain SNPs present in or near genes associated with obesity and diabetes mellitus such as \textit{FADS2} (\textit{Fatty acid desaturase 2}), \textit{LEP} (\textit{Leptin}), \textit{PPAR-\(\gamma\)} (\textit{Peroxisome Proliferator-Activated Receptor gamma}), as well as selenoproteins \textit{SEP15} (\textit{Selenoprotein 15KDa}) and \textit{SEPS1} (\textit{Selenoprotein S}) were a significant risk for PCa along with a \textit{Cytochrome P450 Family 24 Subfamily A Member 1} (\textit{CYP24A1}) involved in the degradation of Vitamin D3. We have also observed that the SNPs that may be vulnerable to environmental conditions may be playing a role in the initiation of non-aggressive PCa, as all the SNPs that were identified as statistically significant lost their power when adjusted for the effect of the environmental factors. These analyses were carried out by using PLINK software version 1.07.

As the progression of PCa was mapped, we found an increasing role of environmental factors interacting with a panel of SNPs in increasing the risk of aggressive PCa. Among these SNPs, several including those on or near the genes \textit{Myeloma Overexpressed} (\textit{MYEOV}), \textit{Microseminoprotein B} (\textit{MSMB}), \textit{Fatty Acid Desaturase 2} (\textit{FADS2}), \textit{Peroxisome proliferator-activated receptor gamma} (\textit{PPAR-\(\gamma\)}), \textit{SEP15} and \textit{Kallikrein-3} (\textit{KLK3}) showed increased risk of aggressive PCa with \textit{MYEOV} and \textit{PPAR-\(\gamma\)} identified to have the highest statistical significance. These results motivated us to consider a wider approach, and work further on the data to identify if there were any other possibilities of finding potential SNP biomarkers missed by us. We, therefore, considered the statistical adjustments for ageing and each lifestyle factor individually and SNP-SNP epistasis as risk for aggressive PCa. With only some of the prior identified statistically significant SNP biomarkers detected after a number of analyses, we proceeded with a machine-learning approach employing Artificial Neural Networks (ANN) produced by Waikato Environment for Knowledge Analysis (WEKA) version 3.8.1 to understand our data better.
Comparing the statistically significant SNP genotyping risk for aggressive PCa in our cohort between the results obtained by using PLINK analysis and ANN, we successfully narrowed down our list of SNPs to only 3 (rs17793693 in the gene *PPAR-γ*, rs10896438 in the gene *MYEOV* and rs10244329 in the gene *LEP*) from 97 (mentioned in the Appendix). This was a major end-point of this exercise.

The role of *MYEOV*- a putative oncogene, *PPAR-γ* (gene involved with fat metabolism) and *Anaplastic Lymphoma Kinase (ALK)* was verified by analysing the expression of these three proteins in three PCa cell lines (and additional non-prostate cancer cell lines). Of interest here was the involvement of the gene *ALK*, which had previously not been studied by us. The gene *ALK* is involved with non-small lung cancer, and since tobacco smoking was identified as an important risk factor, we decided to involve the gene *ALK* as well while plotting the phylogenetic tree and doing protein expression analysis. This work helped us identify a unique link between risk of PCa and the gene *ALK*. We performed a multi-dimensional analysis and these findings were not only interesting, but also very novel in its own way.

Overall, we conclude that multidimensional/holistic approach can be designed to discover the hidden biomarkers in biomedical ailments which are still in want of one. Using this approach successfully, we herewith propose SNP biomarkers in three genes for aggressive PCa—*MYEOV, LEP* and *PPARG* for NZ cohort. By extending the findings of multidimensional approach, we were able to report the over-expression of an oncogene- *MYEOV* and a kinase- *ALK* that were identified to be in close proximity when analysed using phylogenetic analysis in aggressive PCa.
Acknowledgements

This study was carried out at the Faculty of Medical and Health Science (FMHS), University of Auckland (UoA), Auckland during the years 2013-2017.

PhD thesis is a dedicated long-haul effort, which requires countless inputs, guidance and help from many people. Consequently, there are many more who should be stated, but there is insufficient room to mention them all. Thus, I would like to thank very people who are not named below for their boundless co-operation in producing this thesis. Without their involvement, this work, and the period that I have consumed in research and writing, would have been poorer. I am indebted to their priceless contribution.

First and foremost, I would like to express my deepest gratitude to my Main Supervisor, Professor Lynnette R. Ferguson, FMHS, for the opportunity to carry out my PhD research in her research laboratory, constant mentoring and for the critical feedback from time-to-time. Without her guidance, I would have been lost.

I am indebted to my Co-Supervisors, Dr Nishi Karunasinghe, Research Fellow, Auckland Cancer Society Research Centre (ACSRC) and Dr Anower Jabed, Research Fellow, FMHS, and Advisors Dr Gareth Marlow, Research Fellow, Cardiff University and Dr Prasanna Kallingappa, Research Fellow, VJU-FMHS, for their intellectual input, backing and encouragement. This PhD work has been in excess of unbiased education of the scientific proofs, gathering the methodical verdicts and compiling them into the thesis. It has been a hard, but a delightful journey throughout, mainly because of these afore mentioned people. I owe them all a great deal of gratitude and respect for their priceless guidance to achieve this goal. It will be wrong, however, to not make a special mention of Nishi and Prasanna for their dedicated and unbiased support all through this wonderful learning curve.
I have also been blessed with guidance and motivation from Professors, Research Fellows and Tutors from across the University. It will be injustice to not acknowledge the support I received from Professor Trevor Shervin (Associate Dean, FMHS), Associate Professor Clare Wall (Head of the Department, Nutrition), Professor David Cameron-Smith (Chair in Nutrition, Liggins Institute, FMHS), Dr Graeme Finlay (ACSRC, FMHS), Mr Philip Shepherd (Sequenom Facility, Liggins Institute, FMHS), Dr Christina Buchanan (FMHS), Dr Marjan Askarian-Amiri (ACSRC, FMHS), Dr Cherie Blenkiron (FMHS), Dr Jo Perry (Liggins Institute), Dr Selvan Reddy (School of Biological Science, UoA), Dr Jagdish Jaiswal (ACSRC, FMHS) and Dr Karen Bishop (FMHS) and the help in various laboratory-based aspects from Ms Pamela Murray (ACSRC, FMHS). A special mention of Professor Ajit Narayanan, Professor and Head of Research, Auckland University of Technology (AUT), should be made for his valuable input regarding the artificial neural network chapter and guidance all through the course of my PhD.

I would also like to thank my fellow postgraduate students Dr Khanh Tran (FMHS), and Mr Elisha Hayden (FMHS), for helping me out in day-to-day laboratory work and sticking with me to encourage during my lows.

If there is someone whom I owe a great deal in this entire journey are my dearest friends, Mr Vijay Naidu (AUT), Mr Chi Hsiu-Juei Kao (FMHS), and Dr Alice Wang (FMHS). I can’t express enough gratitude for the help, encouragement and assistance I received from them, both professionally and personally. They were always there in all my highs and especially lows. Without them, great deal of this work would not have been easy to materialise. I owe you guys big time.

Since this PhD was pursued without any scholarship support, I would also like to thank all my employers and colleagues for the work provided to me and the support during the various
shifts. Ms Victoria Henderson, Dr Helen Howard-Jones, Mrs Anna Li, Mrs Ivo Syayadi, and Dr Sheryl Munro (Examinations, UoA), and Ms Latifa Khan and Dr Ho Joon Lee (Molecular Medicine and Pathology Labs, FMHS), among others, have been particularly very helpful in this regards and this entire journey would not have been possible without their support.

My family is an integral part of my life. Thanks to the present-day communication tools, I never felt separated from them even with the >10,000 km distance. I am at loss to express my appreciation to my parents, Mrs. Sujata Vaidyanathan (Matéy), Mr. Vaidyanathan Venkateswaran (Dad), my sister and brother-in-law Mrs. Gitanjali Chandel and Mr. Mohit Chandel, and my niece, our little angel- Baby Arya Chandel (Rimi)- who was born during this PhD. I am blessed to have such parents and family who always believed in me and reinforced me in all parts of my life.

Finally, I would like to express my deepest appreciation to my then girlfriend and now wife, Radha Pallati, without whom this journey would have been incomplete, joyless and in-a-way meaningless. She is the source of my inspiration. I thank her for all the hardship she endured and sacrifices she made to support me to reach this stage in my life. She has always bestowed me the encouragement, courage and even ideas to fulfil this PhD thesis.
To my parents
List of Publications and Conference Contributions

Published and unpublished (under peer-review) manuscripts in peer-reviewed journals produced based on the work done in the current project during the candidature including some of the conference presentations by the candidate:

ARTICLES READY TO BE SUBMITTED

Research Article

Artificial Neural Networks: Unravelling “hidden” SNP biomarkers for aggressive prostate cancer - a New Zealand case-control study
Venkatesh Vaidyanathan, Vijay Naidu, Chi Hsiu-Juei Kao, Radha Pallati, Prasanna Kallingappa, Nishi Karunasinghe and Ajit Narayanan

Abstract: Artificial Neural Networks is a crucial machine-learning method employed to define the role of advanced bioinformatics and the identification of certain huge non-processable data for human brain. This method of creaming the undefined role of certain factors on the risk of diseases, such as aggressive prostate cancer is not only unique in that the machine-learning tools are “self-trained” and run with minimal human intervention(s). we have compared the results, herewith, between the association of certain SNPs are risk for aggressive PCa obtained by using established software and with the one developed by artificial intelligence, and have narrowed down our overall number of SNPs associated with risk of aggressive disease considerably.

Research Article

Role of MYEOV-PPARγ-ALK nexus and obesity in the risk of prostate cancer- a New Zealand model
Venkatesh Vaidyanathan, Vijay Naidu, Khanh Tran, Elisha Hayden, Anower Jabed, Prasanna Kallingappa, Chi Hsiu-Juei Kao, Nishi Karunasinghe, Radha Pallati, Graeme Finlay, Pamela Murray, Gareth Marlow, Ajit Narayanan and Lynnette R. Ferguson

Abstract: The gene x environment impact on incidence and progression of diseases is very unique and fascinating, but at the same time generates hope for the many patients suffering with chronic diseases such as prostate cancer (PCa). We have performed a multi-directional study herewith and have identified a very unique and not yet reported link between three genes - MYEOV, ALK, PPAR-γ and the risk of PCa using a case-control study and validation using protein expression in imported cell lines. The result of this unique demographic and molecular analysis on risk of PCa and the potential role of MYEOV, ALK, PPAR-γ is herewith detailed.

ACCEPTED ARTICLE (IN PRESS)

Research Letter

Aggressive prostate cancer incidence in New Zealand- “united we fall, divided we stand” (New Zealand Medical Association Journal, Vol 130 No. 1466, ISSN: 1175-8716- to be online on 01-12-2017. www.nzma.org.nz/journal)
Venkatesh Vaidyanathan, Nishi Karunasinghe, Vetrivhel Krishnamurthy, Chi Hsiu-Juei Kao, Vijay Naidu, Radha Pallati, Alice Wang, Khanh Tran, Prasanna Kallingappa, Anower Jabed, Syed M. Shahid, Jonathan Masters, Clare Wall, Ajit Narayanan, Lynnette R. Ferguson

Abstract: Prostate cancer is an important health burden to the healthcare system of any country. However, with the current prostate-specific antigen biomarker having low predictive value even for diagnostic purposes, the challenge is still open to tackle this chronic disease. There have been a number of studies which have indicated and encouraged a multi-directional approach to combat this
disease. We have been carrying out a multi-directional approach in order to identify certain New Zealand-specific factors which may be drivers for this cancer and its aggressive forms. These will be explained in further detail in this research letter.

**PUBLISHED ARTICLES**

**Research Report**

SNP-SNP interactions as risk factors for aggressive prostate cancer (doi:10.12688/f1000research.11027.1)
Venkatesh Vaidyanathan, Vijay Naidu, Nishi Karunasinghe, Anower Jabeed, Radha Pallati, Gareth Marlow, Lynnette R. Ferguson

**Abstract:** Prostate cancer (PCa) is one of the most significant male health concerns worldwide. Single nucleotide polymorphisms (SNPs) are becoming increasingly strong candidate biomarkers for identifying susceptibility to PCa. We identified a number of SNPs reported in genome-wide association analyses (GWAS) as risk factors for aggressive PCa in various European populations, and then defined SNP-SNP interactions, using PLINK software, with nucleic acid samples from a New Zealand cohort. We used this approach to find a gene x environment marker for aggressive PCa, as although statistically gene x environment interactions can be adjusted for, it is highly impossible in practicality, and thus must be incorporated in the search for a reliable biomarker for PCa. We found two intronic SNPs statistically significantly interacting with each other as a risk for aggressive prostate cancer on being compared to healthy controls in a New Zealand population.

**Research Article**

Effect of ageing and Single Nucleotide Polymorphisms associated with risk of aggressive prostate cancer in a New Zealand population (doi: 10.1039/C7MB00203C)
Venkatesh Vaidyanathan, Vijay Naidu, Nishi Karunasinghe, Chi Hsiu-Juei Kao, Radha Pallati, Anower Jabeed, Gareth Marlow, Prasanna Kallingappa, Lynnette R. Ferguson

**Abstract:** Prostate cancer is one of the most significant male health concerns worldwide, various researchers carrying out molecular diagnostics have indicated that genetic interactions with biological and behavioral factors play an important role in the overall risk and prognosis of this disease. Single nucleotide polymorphisms are increasingly becoming strong biomarker candidates to identify susceptibility of prostate cancer. We carried out risk association of different stages of prostate cancer to a number of single nucleotide polymorphisms to identify the susceptible alleles in a New Zealand population and checked the interaction with environmental factors as well. We have identified a number of single nucleotide polymorphisms to have associations specifically to the risk of prostate cancer and aggressiveness of the disease, and also certain single nucleotide polymorphisms to be vulnerable to the reported behavioral factors.

**Review**

Prostate Cancer: Is It a Battle Lost to Age? (doi: 10.3390/geriatrics1040027)
Venkatesh Vaidyanathan, Nishi Karunasinghe, Anower Jabeed, Radha Pallati, Chi Hsiu-Juei Kao, Alice Wang, Gareth Marlow and Lynnette R. Ferguson

**Abstract:** Age is often considered an important non-modifiable risk factor for a number of diseases, including prostate cancer. Some prominent risk factors of prostate cancer include familial history, ethnicity and age. In this review, various genetic and physiological characteristics affected due to advancing age will be analyzed and correlated with their direct effect on prostate cancer.

**Research Article**

Abstract: Prostate cancer is one of the most significant male health concerns worldwide. Numerous researchers carrying out molecular diagnostics have indicated that genetic interactions with biological and behavioral factors play an important role in the overall risk and prognosis of this disease. Single nucleotide polymorphisms (SNPs) are increasingly becoming strong biomarker candidates to identify susceptibility to prostate cancer. We carried out a gene x environment interaction analysis linked to aggressive and non-aggressive prostate cancer (PCa) with a number of SNPs. By this method, we identified the susceptible alleles in a New Zealand population, and examined the interaction with environmental factors. We have identified a number of SNPs that have risk associations both with and without environmental interaction. These indicate that certain SNPs have been associated with disease vulnerability based on behavioral factors.

(Shortlisted as “2017 Hot Articles in Molecular Biosystems” and appeared on the Front Cover of the Issue)

POSTERS PRESENTED AND TALKS DELIVERED

SNP genotyping and risk of aggressive prostate cancer- seeing beyond what GWAS ever taught us

Understanding SNP genotyping data- are we there yet?
Vaidyanathan Venkatesh.

Are we eating our way to aggressive prostate cancer: a case-control study?
Vaidyanathan Venkatesh.
(Shortlisted for three-minute Thesis Presentation Award)

New Zealand lifestyle, increased BMI and risk of aggressive Prostate Cancer (doi: 10.7490/f100research.1114377.1)
Vaidyanathan Venkatesh, Karunasinghe Nishi, Masters Jonathan, Ferguson Lynnette.
(Shortlisted for Best Poster Award)

Lifestyle, increased BMI and aggressive Prostate Cancer- from bench to table (doi: 10.13140/RG.2.2.27762.30405)
Vaidyanathan Venkatesh, Karunasinghe Nishi, Naidu Vijay, Jaber Anower, Ferguson Lynnette.
New Zealand Society for Oncology Conference, Palmerston North. New Zealand, October 2016.
(Shortlisted for Young Researcher Award; Awarded Travel Grants by NZSO and FMHS-PGSA)

Prostate cancer, aggressiveness and red meat consumption-a Nutrigenomics perspective (doi: 10.13140/RG.2.1.4478.5522)
Vaidyanathan Venkatesh, Bishop Karen, Karunasinghe Nishi, Han Dug Yeo, Zhu Shuotun, Masters Jonathan, Ferguson Lynnette.
New Zealand Society for Oncology Conference, Tauranga. New Zealand, October 2014.
Table of Contents

Abstract ................................................................................................................................................... ii
Acknowledgements ................................................................................................................................. v
List of Publications and Conference Contributions ................................................................................ ix
Table of Contents .................................................................................................................................. xii
List of Figures ........................................................................................................................................ xix
List of Tables .......................................................................................................................................... xx
List of Abbreviations ............................................................................................................................ xxii
Chapter 1. Introduction and Background .......................................................................................... 2
Chapter 2. Review of Literature .............................................................................................................. 6
  2.1. Prostate Cancer ............................................................................................................................ 6
    2.1.1 Symptoms and implications of PCa ....................................................................................... 7
    2.1.2 Screening and diagnosis ........................................................................................................ 7
    2.1.3 Staging, grading and risk stratification ................................................................................. 8
    2.1.4 Risk factors of PCa .............................................................................................................. 11
  2.2. Prostate Cancer and lifestyle factors ........................................................................................... 13
    2.2.1. Smoking status .................................................................................................................... 13
    2.2.2. Energy consumption and Body Mass Index ........................................................................ 15
    2.2.3. Alcohol consumption ........................................................................................................ 17
    2.2.4. Dietary factors ..................................................................................................................... 18
  2.3. Prostate Cancer and Genetic factors ........................................................................................... 21
    2.3.1. Genome Wide Association Studies ..................................................................................... 21
    2.3.2. Single Nucleotide Polymorphism Genotyping ................................................................. 22
    2.3.3. Methods for detecting SNPs ............................................................................................. 23
    2.3.4. Statistical analysis ............................................................................................................... 24
      2.3.4.1. PLINK based Identity-by-state (IBS) clustering.......................................................... 25
      2.3.4.2. Artificial Neural Networks ........................................................................................... 25
  2.4. Chapter Summary ...................................................................................................................... 27
Chapter 3. Hypothesis, Aims and Plot of the Thesis ......................................................................... 29
  3.1. Rationale of the Study ................................................................................................................. 29
  3.2. Limitations of the Study ............................................................................................................. 29
  3.3. Aims of the thesis ....................................................................................................................... 33
3.4. Thesis Layout .............................................................................................................................. 34

Chapter 4. Materials and Methods ....................................................................................................... 37

4.1. Patient recruitment for the study .............................................................................................. 38
4.2. Collection of human samples: .................................................................................................... 39
4.3. SNP Selection: .............................................................................................................................. 40
4.4. SNP Genotyping using iPLEX MassARRAY technology ............................................................... 42
   4.4.1. Principle .................................................................................................................................. 42
   4.4.1.1. DNA extraction using QIAcube™ ................................................................................. 42
   4.4.1.2. iPLEX MassARRAY ......................................................................................................... 42
   4.4.2. Reagents and Equipments ................................................................................................... 44
   4.4.3. Protocol ............................................................................................................................. 44
   4.4.3.1. DNA Extraction from EDTA blood ................................................................................. 44
   4.4.3.2. iPLEX preparation ......................................................................................................... 44
   4.4.3.3. DNA plating .................................................................................................................. 45
   4.4.3.4. PCR reaction ................................................................................................................. 45
   4.4.3.5. SAP Treatment ............................................................................................................. 46
   4.4.3.6. Extension primer optimisation .................................................................................... 47
   4.4.3.7. iPLEX Extension ............................................................................................................ 48
   4.4.3.8. Resin clean-up .............................................................................................................. 49
   4.4.3.9. Data analysis ................................................................................................................ 50
4.5. Statistical Analysis ...................................................................................................................... 52
   4.5.1. Lifestyle and demographics analyses ................................................................................. 52
   4.5.2. SNP genotyping Data Analysis using PLINK ......................................................................... 52
   4.5.2.1. Data analysis ................................................................................................................ 52
   4.5.2.2. Association analysis ..................................................................................................... 53
   4.5.2.3. Interaction analysis ...................................................................................................... 54
   4.5.2.4. Epistasis analysis .......................................................................................................... 56
   4.5.3. SNP genotyping Data Analysis using Artificial Neural Networks ......................................... 56
   4.5.3.1. WEKA ............................................................................................................................ 56
   4.5.3.2. ANN ............................................................................................................................. 57
   4.5.3.3. MLP ............................................................................................................................. 57
   4.5.3.4. ARFF File Creation ........................................................................................................ 59
   4.5.3.5. First Process of MLP ..................................................................................................... 59
   4.5.3.6. Data Analysis – Stage II: Weights (SNPs) Extraction .................................................... 60
4.5.3.7. Data Analysis – Stage III: Second MLP Process ............................................................. 61
4.5.3.8. ARFF File Creation ........................................................................................................ 61
4.5.3.9. Second Process of MLP ................................................................................................ 62

4.6. Cell line studies ...................................................................................................................... 63
4.6.1. Reagents and Equipment .................................................................................................... 63
4.6.2. Protocol ............................................................................................................................... 64
4.6.2.1. Thawing cell lines ......................................................................................................... 64
4.6.2.2. Subculture procedure .................................................................................................. 65
4.6.2.3. Cryopreservation of cells ............................................................................................. 65

4.7. Western Blot .............................................................................................................................. 66
4.7.1. Introduction ....................................................................................................................... 66
4.7.2. Reagents and Equipment .................................................................................................. 66
4.7.3. Protocol ............................................................................................................................... 68
4.7.3.1. Protein lysate preparation (taking a 6-well plate as an example) ............................... 68
4.7.3.2. Protein quantification .................................................................................................. 68
4.7.3.3. Running PAGE gel and transferring .............................................................................. 68
4.7.3.4. Antibody blotting ......................................................................................................... 69
4.7.3.5. Stripping and re-probe ................................................................................................. 69

Chapter 5. SNP association with risk of prostate cancer ............................................................ 71

5.1. Environmental factors and risk of aggressive prostate cancer among a population of New Zealand men- a genome-wide association analysis ....................................................... 73
5.1.1. Introduction ....................................................................................................................... 74
5.1.2. Materials and Methods ................................................................................................... 75
5.1.2.1. Study population .......................................................................................................... 75
5.1.2.2. Data collection for demography, and lifestyle details ................................................. 76
5.1.2.3. Collection and processing of blood samples ............................................................... 76
5.1.2.4. Selection of SNPs ......................................................................................................... 77
5.1.2.5. Genotyping of candidate genes ................................................................................... 78
5.1.2.5.a. SNP genotyping by Sequenom multiplexing ............................................................. 78
5.1.2.5.b. SNP genotyping by TaqMan® assay ........................................................................ 78

5.1.3. Collection of clinical characteristics of patients ................................................................. 79
5.1.3.1. Definition of aggressiveness ........................................................................................ 79
5.1.3.2. Statistical analysis ........................................................................................................ 79
5.1.3.2.a. SNP data cleaning....................................................................................................... 79
5.1.3.2.b. Gene-environment interaction for risk of PCa ................................................................. 80
5.1.3.2.c. Significance of demographic factors ............................................................................. 81
5.1.3.2.d. SNP association analysis for risk of Prostate Cancer .................................................. 81

5.1.4. Results .................................................................................................................................. 81
5.1.4.1. Gene x -environment effects and prostate cancer risk in a New Zealand population 81
5.1.4.2. Pathology, BMI and lifestyle .......................................................................................... 84
5.1.4.3. Correction for the effect of covariates to identify the genic risk of aggressive PCa ... 87

5.1.5. Discussion ............................................................................................................................. 88
5.1.5.1. Effect of environment and SNP genotype with risk of prostate cancer ...................... 89

5.1.5.1.a. SNP genotype analysis of non-aggressive prostate cancer vs healthy controls 90
5.1.5.1.b. SNP genotype analysis of aggressive prostate cancer vs non-aggressive prostate cancer ......................................................................................................................... 92
5.1.5.1.c. SNP genotype analysis of aggressive prostate cancer vs healthy controls ...... 93

5.1.5.2. Age and risk of prostate cancer ....................................................................................... 97
5.1.5.3. BMI, smoking tobacco, and alcohol consumption at recruitment and risk of prostate cancer .............................................................................................................................. 97
5.1.5.4. Correction for gene x environment interaction and risk of prostate cancer ............... 99

5.1.5.4.a. Genotype and the outcome of non-aggressive prostate cancer (vs healthy controls) ................................................................................................................................. 99
5.1.5.4.b. Genotype and the outcome of aggressive prostate cancer (vs non-aggressive prostate cancer) .................................................................................................................. 101
5.1.5.4.c. Genotype and the outcome of aggressive prostate cancer (vs healthy controls) 102

5.1.6. Conclusions .......................................................................................................................... 105

5.2. Section summary and preamble to next section ..................................................................... 108

5.3. Prostate Cancer: Is It a Battle Lost to Age? .......................................................................... 110
5.3.1. Introduction .......................................................................................................................... 110

5.3.1.1. Risk Factors for Prostate Cancer .................................................................................... 112
5.3.1.2. Aging, Immunity, Inflammatory Response, and Prostate Cancer .............................. 113
5.3.1.3. Aging, Cholesterol Metabolism, and Prostate Cancer ................................................ 116
5.3.1.4. Aging, Testosterone Level and Prostate Cancer .......................................................... 118
5.3.1.5. Aging, Genetic and Epigenetic Effects and Prostate Cancer ....................................... 122
5.3.2. Discussion .......................................................................................................................... 125

5.4. Effect of ageing and Single Nucleotide Polymorphisms associated with risk of aggressive prostate cancer in a New Zealand population ........................................................................... 129
6.1.2. Methods ............................................................................................................................................. 169
6.1.2.1. Selection of Data Set ......................................................................................................................... 169
6.1.2.2. Data tidying and pre-processing ........................................................................................................ 169
6.1.2.3. SNP Genotype analysis using PLINK software .................................................................................. 169
6.1.2.4. Data Analysis—Stage I: First MLP Process ..................................................................................... 170
  6.1.2.4.a. WEKA ........................................................................................................................................... 170
  6.1.2.4.b. ANN ............................................................................................................................................... 171
  6.1.2.4.c. Multi Layer Perceptron .............................................................................................................. 171
  6.1.2.4.d. ARFF File Creation ..................................................................................................................... 173
  6.1.2.4.e. First Process of MLP .................................................................................................................. 173
6.1.2.5. Data Analysis—Stage II: Weights (SNPs) Extraction ........................................................................ 174
6.1.2.6. Data Analysis—Stage III: Second MLP Process .............................................................................. 174
  6.1.2.6.a. ARFF File Creation ..................................................................................................................... 175
  6.1.2.6.b. Second Process of MLP ........................................................................................................... 175
6.1.3. Results ............................................................................................................................................... 175
6.1.4. Discussion ......................................................................................................................................... 183
6.2. Chapter summary ................................................................................................................................. 186

Chapter 7. Protein Expression Analysis ...................................................................................................... 188

7.1. Role of MYEOV-PPARY-ALK nexus and tobacco smoking and obesity in the risk of prostate cancer- a New Zealand model .............................................................................................................. 189

7.1.1. Introduction ........................................................................................................................................ 189
7.1.2. Materials and Methods ....................................................................................................................... 191
  7.1.2.1. Statistical analysis of demographics factor and SNP genotype data (association and interaction): ................................................................. 191
  7.1.2.2. Phylogenetic tree construction: ...................................................................................................... 193
  7.1.2.3. Western blot analyses: .................................................................................................................. 194
      7.1.2.3.a. Cell lines: .................................................................................................................................. 194
      7.1.2.3.b. Western blot assay: .................................................................................................................. 194
      7.1.2.3.c. Quantification of protein: ......................................................................................................... 195
7.1.3. Results and Discussion ......................................................................................................................... 195
  7.1.3.1. BMI and pathology: ....................................................................................................................... 195
  7.1.3.2. Genetic polymorphism variations and risk of prostate cancer: .................................................... 196
  7.1.3.3. Cladistics-based phylogenetic tree: ............................................................................................... 199
  7.1.3.4. Protein expression levels: ............................................................................................................. 204
7.1.4. Conclusions ....................................................................................................................... 207
Chapter 8. Final Discussion ........................................................................................................ 210
Chapter 9. Future Directions and application of work.............................................................. 215
  9.1. Aggressive prostate cancer incidence in New Zealand- ‘united we fall, divided we stand’ .... 216
Chapter 10. Bibliography ........................................................................................................ 221
Chapter 11. Appendices ........................................................................................................... 252
  11.1. SUPPLEMENTARY TABLE ................................................................................................. 252
  11.2. SUPPLEMENTARY IMAGES OF Typer 4.0 MAPS ........................................................ 267
  11.3. SUPPLEMENTARY FIGURE- LD Map ............................................................................ 272
  11.4. Co-authorship Forms ..................................................................................................... 273
List of Figures

Figure 3.1 PhD Thesis Flow Chart ................................................................. 35

Figure 4.1 A typical Call Cluster Plot. The call for a sample (well no. 1A) is called for here for the SNP rs6162. ................................................................. 51

Figure 5.1 Pictorial Abstract ........................................................................ 73

Figure 5.2 Individual aggressive PCa risk association with SNPs, environmental factors as well as in combination of both compared to non-aggressive disease. ........................................ 100

Figure 5.3 Individual aggressive PCa risk association with SNPs, environmental factors as well as in combination of both compared to controls ................................................................. 104

Figure 5.4 Pictorial conclusion illustrating the various associations and links leading to aggressive and non-aggressive PCa. ......................................................... 106

Figure 5.5 Relation between aging and risk and progression of prostate cancer. ........................................ 112

Figure 5.6 Relation between aging and immunity ............................................................................. 114

Figure 5.7 Relation between cholesterol metabolism and progression of prostate cancer. .................. 118

Figure 5.8 Relation between aging, cholesterol metabolism and hormonal imbalance. ..................... 121

Figure 5.9 Various pathways and the genes identified to be significantly associated with a risk of aggressive prostate cancer (compared to healthy controls). ............................................ 148

Figure 5.10 Various pathways and the genes identified to be significantly associated with a risk of aggressive prostate cancer (compared to non-aggressive prostate cancer) .... 152

Figure 6.1: Three layers connectivity with neuron structure. .......................................................... 167

Figure 6.2: Schematic representation of the methodology ................................................................. 168

Figure 6.3: SNPs significantly associated with risk of aggressive PCa using IBS and ANN analyses. .. 184

Figure 7.1 Cladistics-based phylogenetic tree. ................................................................. 201

Figure 7.2 Western Blot for the expression level of proteins .............................................................. 205

Figure 7.3 Quantification of expression levels of ALK across various cell lines. .............................. 205

Figure 7.4 Quantification of expression levels of PPAR-γ across various cell lines. ......................... 206

Figure 7.5 Quantification of expression levels of MYEOV across various cell lines. ....................... 207
List of Tables

Table 2.1: TNM staging system for PCa (Sobin and Fleming 1997) ........................................................ 9
Table 2.2: Gleason patterns (Gleason 1992) ............................................................................................ 10
Table 3.1: Power calculation for age as a parameter for prostate cancer vs healthy controls .......... 30
Table 3.2: Power calculation for BMI as a parameter for prostate cancer vs healthy controls .......... 31
Table 4.1 Reagents ................................................................................................................................ 44
Table 4.2 PCR Master Mix ........................................................................................................................ 46
Table 4.3 PCR Programme ....................................................................................................................... 46
Table 4.4 SAP Master Mix ........................................................................................................................ 47
Table 4.5 SAP Treatment Programme .................................................................................................. 47
Table 4.6 Division of the groups of SNPs based on their mass ............................................................. 48
Table 4.7 Extension Master Mix ............................................................................................................ 49
Table 4.8 Extension PCR Programme .................................................................................................... 49
Table 4.9 Composition of media for cell culture .................................................................................. 63
Table 4.10 Phosphate Buffered Saline (PBS) ......................................................................................... 64
Table 4.11 Equipment ........................................................................................................................... 64
Table 5.1 Statistically significant case-control SNP association between patients with non-aggressive prostate cancer and healthy controls ......................................................................................... 82
Table 5.2 Statistically significant case-case SNP association between patients with aggressive prostate cancer and non-aggressive prostate cancer ........................................................................... 83
Table 5.3 Statistically significant case-control SNP association between patients with aggressive prostate cancer and healthy controls ................................................................................................... 83
Table 5.4 The association between BMI and risk and/or aggressiveness of prostate cancer .............. 85
Table 5.5 Association between tobacco smoking status and risk and/or aggressiveness of prostate cancer .................................................................................................................................................. 86
Table 5.6 Comparison of alcohol consumption and risk and/or aggressiveness of prostate cancer ... 86
Table 5.7 Logistic model of SNPs risk for aggressive prostate cancer vs non-aggressive prostate cancer after correcting for interaction with multiple covariates (BMI, tobacco smoking, and alcohol consumption) ........................................................................................................................................ 88
Table 5.8 Logistic model of SNPs risk for aggressive prostate cancer vs healthy controls after correcting for interaction with multiple covariates (BMI, tobacco smoking, and alcohol consumption) ........................................................................................................................................ 88
Table 5.9 Association between age and aggressive prostate cancer vs healthy controls ................. 132
Table 5.10 Association between age and aggressive prostate cancer vs non-aggressive prostate cancer ................................................................................................................................................. 132

Table 5.11 Association between age and non-aggressive prostate cancer vs healthy controls. ...... 133

Table 5.12 Statistically significant SNP associated with gene x environment effect on risk of aggressive prostate cancer v/s healthy controls after adjusting for each environmental parameter individually and along with age .......................................................................................................... 134

Table 5.13 Statistically significant SNP associated with gene x environment effect on risk of aggressive prostate cancer v/s non-aggressive prostate cancer after adjusting for each environmental parameter individually and along with age ........................................................................................................ 137

Table 5.14 Statistically significant SNP associated with gene x environment effect on risk of non-aggressive prostate cancer v/s healthy controls after adjusting for each environmental parameter individually and along with age ........................................................................................................ 139

Table 5.15 New Zealand factors” and risk of non-aggressive prostate cancer ................................ 154

Table 5.16 Statistically significant SNP-SNP interactions discovered in patients with aggressive PCa when compared to healthy controls............................................................... 159

Table 6.1: Results identifying risk of SNPs for PCa using IBS Clustering and ANN ......................... 176

Table 6.2: Results identifying risk of SNPs for aggressive PCa using IBS Clustering and ANN .......... 177

Table 6.3: Results of the first process of MLP (Stage-I) for the first and second analysis .......... 180

Table 6.4 Results of the second process of MLP (Stage-III) for the first and second analysis .......... 182

Table 7.1: Association between high BMI status and the risk of prostate cancer ......................... 196

Table 7.2: Association between tobacco smoking status and the risk of prostate cancer ............. 196

Table 7.3 Statistically significant SNP association between patients with prostate cancer and healthy controls before and after adjusting for BMI ........................................................................................................ 197
List of Abbreviations

ABI  Applied Biosystems
ADD  additive
ADT  androgen deprivation treatment
AKR1C3  Aldo-keto reductase family 1 member C3
ALK  Anaplastic Lymphoma Kinase
ANXA1  Annexin A1
BH-FDR  Benjamini and Hochberg false discovery rate
BLOSUM  BLOcks SUbstitution Matrix
BMI  body mass index
BONF  Bonferroni correction(s)
cAMP  cyclic adenosine monophosphate
CCHCR1  Coiled-coil alpha-helical rod protein 1
CRISPR  clustered regularly interspaced short palindromic repeats
CRISPR/Cas9  CRISPR associated protein 9
CTBP2  C-terminal-binding protein 2
DASH  Dynamic allele-specific hybridization
df  degree of freedom
DHEA  dehydroepiandrosterone
DNA  deoxyribonucleic acid
dNDPs  deoxynucleotide diphosphates
dNTPs  deoxynucleotide triphosphates
DOMDEV  dominant deviation
EDTA  Ethylene Diamine Tetra Acetic acid
ER  endoplasmic reticulum
FADS2  Fatty acid desaturase 2
FASN  Fatty Acid Synthase
FASTA  FAST-ALL
FYCO1  (FYVE and coiled-coil domain containing 1)
GDF15  Growth differentiation factor 15
gDNA  genomic DNA
GeCKO  Genome-Scale CRISPR Knock-Out
GENO_2DF  general_2 df joint test of both additive and dominant deviation
GRN  Granulin
GST1  Glutathione S-Transferase-1
GSTP1  Glutathione S-Transferase P
GWAA  Genome-wide association analysis
GWAS  Genome-wide association studies
HDL  high-density lipoprotein
HDR  homology directed repair
HFDs  high fat diets
HNF1B Hepatocyte Nuclear Factor 1 homeobox B
HOXB8 Homeobox B8
HWE Hardy Weinberg Equilibrium
IBS identical-by-state
IL10 Interleukin -10
IL1RN Interleukin 1 receptor antagonist
IL8 Interleukin -8
KLK3 Kallikrein-3
KLKP1 Kallikrein pseudogene 1
LD linkage disequilibrium
LDL low-density lipoprotein
LEP Leptin
lncRNAs long non-coding RNAs
MAFFT Multiple Alignment using Fast Fourier Transform
ME Minimum evolution
MIC-1 Macrophage inhibitory cytokine-1
miRNAs microRNAs
ML Maximum likelihood
MLH1 MutL homolog 1
MLP Multi Layer Perceptron
MLPH Melanophilin
MMP9 Matrix metallopeptidase 9
MP Maximum parsimony
mRNA messenger-ribonucleic acid
MSMB Microseminoprotein Beta
MSR1 Macrophage Scavenger Receptor 1
MYEOV Myeloma Overexpressed
ng/μl nanogram per microliter
NHER non-homologous end joining
NJ Neighbour-joining
NSCLC non-small cell lung cancer
NUDT11 Nucleoside Diphosphate-linked Moiety X Motif 11
PAM Point Accepted Mutations
PCA prostate cancer
PODXL Podocalyxin-like
PVDF Polyvinylidene difluoride
PPAR-γ Peroxisome proliferator-activated receptor gamma
PSA prostate-specific antigen
RFLP Restriction fragment length polymorphism
RNA ribonucleic acid
RNAi RNA interference
RNAs ribonucleic acids
RNASEL RNase L
ROS reactive oxygen species
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEP15</td>
<td>Selenoprotein 15KDa</td>
</tr>
<tr>
<td>SEPS1</td>
<td>Selenoprotein S</td>
</tr>
<tr>
<td>sgRNA</td>
<td>single guide RNA</td>
</tr>
<tr>
<td>shRNA</td>
<td>short hairpin RNA</td>
</tr>
<tr>
<td>SLC26A6</td>
<td>Solute carrier family 26 member 6</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SRD5A2</td>
<td>Steroid 5α-reductase type 2</td>
</tr>
<tr>
<td>TALEN</td>
<td>Transcription Activator-Like Effector Nuclease</td>
</tr>
<tr>
<td>TERT</td>
<td>Telomerase reverse transcriptase</td>
</tr>
<tr>
<td>TLR4</td>
<td>Toll-like receptor 4</td>
</tr>
<tr>
<td>UPGMA</td>
<td>Unweighted pair group method with arithmetic means</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>VDR</td>
<td>Vitamin D (1,25-dihydroxyvitamin D3) receptor</td>
</tr>
<tr>
<td>ZFN</td>
<td>Zinc Finger Nuclease</td>
</tr>
</tbody>
</table>
CHAPTER ONE

INTRODUCTION

“We cannot start over, but we can begin now, and make a new ending.”

- Zig Ziglar
Chapter 1. Introduction and Background

The first case of prostate cancer (PCa) was reported in 1853 by a surgeon at The London Hospital, J. Adams, by histological examination (Denmeade and Isaacs 2002). That year, New Zealand’s population is estimated to be 61,850 Māori and 29,600 non-Māori (NZ Retrieved 16-04-2017), and chronic ailments such as PCa were unheard of. Today, the population of New Zealand is close to 4.78 million, which is roughly 0.07% of world population; however, currently, New Zealand records more than 3000 new diagnosed cases of PCa and around 600 deaths in an average year (2015). This establishes PCa as one of the most commonly diagnosed cancers and the second most common cause of cancer-related death among men in Australia/New Zealand (IARC, 2008; 2015). New Zealand also records higher PCa related mortality rates compared to that of USA (Aye, Elwood et al. 2014; Ministry of Health 2016).

It is well established that phenotype (diseases condition) is defined as the interaction of genotype plus environment. A targeted approach is herewith discussed to identify the genetic biomarkers for aggressive PCa by employing novel and stringent methods.

Familial history is a strong indicator of risk of PCa too, like many other cancers. Although with an estimated heritability of 40% (Lichtenstein, Holm et al. 2000), and hereditary undoubtedly being a prominent risk for this disease (Bratt 2002); yet various studies have also indicated that genetic interactions with biological and behavioral factors may also play an important role in the overall risk and prognosis of PCa (Schaid 2004; Karunasinghe, Han et al. 2012; Karunasinghe, Lange et al. 2013).

The influence of high body mass index (BMI) and obesity on the risk of PCa, like various other cancers (Calle and Thun 2004; Basen-Engquist and Chang 2011), is perhaps also one of the best examples of genetic and environmental influences on our health in modern times. Various studies, including the ones reported by us (Vaidyanathan, Naidu et al. 2017), have
shown several lifestyle and external factors, such as high BMI, tobacco smoking, alcohol consumption, and progressing age (Vaidyanathan, Karunasinghe et al. 2016) to be associated with an increased risk of PCa (Lund Nilsen, Johnsen et al. 2000; Kogelberg, Frenkel et al. 2002; Nakamura, Kuroki et al. 2009). Variations in the genome are a major contributor to the differences in disease susceptibility amongst individuals (Tweedary and Belmont 2009). Single nucleotide polymorphisms (SNPs) are the most commonly identified variations in a genome, and are also one of the most commonly studied potential biomarkers for various diseases (Perkel 2008).

Certain modern molecular biology analyses employing GWAS claim to have identified various SNPs that are associated with the risk of PCa, risk of aggressive PCa, age of diagnosis, and/or even PCa mortality (Ciampa, Yeager et al. 2011; Liu, Wang et al. 2011; Pomerantz, Werner et al. 2011). However, some SNPs still largely remain uncharacterized, as they are located within the gene-poor regions such as introns and/or are in the intergenic regions (Lu, Zhang et al. 2011). Such SNPs present in the introns and/or intergenic regions can also be used as indicators of profound changes in the expression of certain genes harboring these SNPs (Vaidyanathan, Naidu et al. 2017). Undoubtedly, further analysis of functional SNPs will potentially allow for a better understanding of the complex genetic pathways involved in PCa incidence and progression, but ignoring the influence of intronic and intergenic SNPs and the role of external parameters in the function of these genes harboring such SNPs is not in the best interest of molecular diagnostics.

The aims of this thesis were to analyse the genome x environment interactions, with regards PCa and aggressive PCa, in a case-control study with men of European ethnicity (self-reported) from New Zealand, assess the importance of individual environmental effects by using multi-dimensional bioinformatics tools and finally consider the effects of resulting
relevant SNPs by protein expression of the genes of interest in various prostate cancer cell lines.
“It is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence.”

- William K. Clifford
Chapter 2. Review of Literature

2.1. Prostate Cancer

PCa is a common but complex disease (Vaidyanathan, Naidu et al. 2017). PCa is the second-most commonly diagnosed cancer in men worldwide (Torre, Bray et al. 2015). In the year 2012, an estimated 1.1 million men worldwide were diagnosed to be suffering from PCa (Ferlay, Soerjomataram et al. 2015). Incidence rates of PCa vary significantly across the world (Torre, Bray et al. 2015), however, in New Zealand, it is the most frequently diagnosed cancer among men (Ministry of Health 2016).

Among the developed countries, PCa has been identified most frequently in countries where people follow a Western-style diet, including countries in North America, Western and Northern Europe, and Oceania (IARC, 2008; Bishop, Erdrich et al. 2015). PCa is the fifth leading cause of cancer death in men worldwide with the highest mortality rate recorded in the Caribbean and Southern and Middle Africa, and lowest in Asia (Torre, Bray et al. 2015). Mortality rates for PCa have been decreasing in many developed countries, including in New Zealand, Australia, Canada, the United States, the United Kingdom, Austria, Spain and Switzerland, mainly due to improvements in treatments and/or early detection of the disease (Center, Jemal et al. 2012).

In New Zealand, PCa is the most commonly registered male cancer and the third most common cause of cancer deaths in men, accounting for 27.2% of all male cancer registrations and 13.4% of male cancer deaths in 2013 (Ministry of Health 2016). Mortality rates due to PCa are also higher in NZ than that of United States of America (USA) (SEER Cancer Statistics Review 1975-2012; Ministry-of-Health 2014). According to the Surveillance, Epidemiology and End Results (SEER Cancer Statistics Review 1975-2012) Program Cancer Statistics Review for the US between 1975-2012, mortality rates (adjusted to the IARC world standard population) of PCa were 8.5 and 21 for white and black men respectively (SEER
Cancer Statistics Review 1975-2012). In NZ, mortality rates (standardised to the World Health Organization standard population) of PCa for 2011 were 16.6 and 22.1 for non-Māori and Māori men respectively (Ministry-of-Health 2014). These statistics indicate that in NZ, PCa related mortality rate for non-Māori men is almost double the comparable figures for white men from the USA while the rates for Māori men are comparable with the US black population.

2.1.1 Symptoms and implications of PCa

Common PCa symptoms include the need to urinate frequently along with increased urination during the night (nocturia), sudden urges to urinate, difficulty in maintaining a steady stream of urine, blood in the urine (hematuria) and pain during urination (dysuria) (Hodgson, Obertova et al. 2012). In most cases, PCa is identified as a slow growing chronic disease such that the symptoms may not occur for many years (You, Knudsen et al. 2016). Such are the traits of this disease that there may not be any symptoms in the early stages of PCa (You, Knudsen et al. 2016), and even when the symptoms are evident, they may be similar to the non-malignant conditions (Denmeade and Isaacs 2002). Most patients with localised or low-grade form, or non-aggressive PCa usually survive for ≥10 years unlike the patients with aggressive or high-risk PCa (Cooperberg, Lubeck et al. 2004; Saad 2006; Kent, Penson et al. 2016).

Not all patients develop aggressive PCa, but there is no biomarker yet to define and/or differentiate between it and the non-aggressive PCa.

2.1.2 Screening and diagnosis

Primarily, two main tests are used to assess prostate health - digital rectal examination (DRE), and/or determination of serum prostate-specific antigen (PSA) level (Hoffman 2011; Dotto 2014). Elevations of serum PSA- a glycoprotein produced by the epithelial cells of the
prostate gland, indicates an abnormality in it (Hoffman 2011). This may however, be due to the benign enlargement, inflammation and/or even PCa (Drudge-Coates and Turner 2012). Owing to this, the lower range PSA test results cannot be employed to differentiate between any benign prostate disease and cancer.

Since DRE is performed by physicians feeling the prostate gland, and is individual and location-based, it has its own limitations (Thompson, Thrasher et al. 2007) including the fact that stage I cancers are non-palpable. Moreover, DRE can be used to detect tumours present only in the posterior and lateral parts of the prostate gland. These are some of the major hurdles in the detection of PCa (Epstein 2002).

With the advancement of science, certain new biomarkers and advanced techniques especially the multiparametric magnetic resonance imaging (mpMRI) has been adopted as an important tool to improve the predictive value of PCa screening (Thompson, Moses et al. 2014). Usually, men considered to be in the high risk-category (owing to familial history) and/or with an abnormal DRE and/or high PSA level, and/or with significant PCa suspected through mpMRI, are subjected to prostate biopsies.

2.1.3 Staging, grading and risk stratification

Once an individual is diagnosed with PCa, the tumour must be staged to determine if it is localized or has already spread beyond the prostate gland. An as accurate staging of PCa is crucial for the proper selection of treatment options. The most common classification of cancer stage is the TNM Classification of Malignant Tumours adopted by the American Joint Committee on Cancer (AJCC) (Hutterer and Karakiewicz 2007; AJCC 2009; Cowherd 2012). The TNM staging system is a combination of size and properties of the primary tumour (T), involvement of lymph node (N), and distant metastasis (M) (Table 2.1) (Edge and Compton 2010).
Table 2.1: TNM staging system for PCa (Sobin and Fleming 1997)

<table>
<thead>
<tr>
<th>Evaluation of the primary tumour (T)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TX</td>
<td>Cannot evaluate the primary tumour</td>
</tr>
<tr>
<td>T0</td>
<td>no evidence of tumour</td>
</tr>
<tr>
<td>T1</td>
<td>tumour present, but not detectable clinically or with imaging</td>
</tr>
<tr>
<td>T1a</td>
<td>tumour incidentally found in &lt;5% of PCa tissue resected</td>
</tr>
<tr>
<td>T1b</td>
<td>tumour incidentally found in &gt;5% of PCa tissue resected</td>
</tr>
<tr>
<td>T1c</td>
<td>tumour found in needle biopsy performed due to serum PSA</td>
</tr>
<tr>
<td>T2</td>
<td>tumour can be felt on examination, but not spread outside prostate</td>
</tr>
<tr>
<td>T2a</td>
<td>tumour is in half or less than half of one of the prostate glands’ two lobes</td>
</tr>
<tr>
<td>T2b</td>
<td>tumour is in more than half of one lobe, but not both</td>
</tr>
<tr>
<td>T2c</td>
<td>tumour in both lobes</td>
</tr>
<tr>
<td>T3</td>
<td>tumour spread through the prostate capsule</td>
</tr>
<tr>
<td>T3a</td>
<td>tumour has spread through the capsule on one or both sides</td>
</tr>
<tr>
<td>T3b</td>
<td>tumour has invaded one or both seminal vesicles</td>
</tr>
<tr>
<td>T4</td>
<td>tumour has invaded other nearby structures</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluation of the regional lymph nodes (N)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NX</td>
<td>cannot evaluate the regional lymph nodes</td>
</tr>
<tr>
<td>N0</td>
<td>there has been no spread to the regional lymph nodes</td>
</tr>
<tr>
<td>N1</td>
<td>there has been spread to the regional lymph nodes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluation of distant metastasis (M)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MX</td>
<td>cannot evaluate distant metastasis</td>
</tr>
<tr>
<td>M0</td>
<td>no distant metastasis</td>
</tr>
<tr>
<td>M1</td>
<td>distant metastasis</td>
</tr>
<tr>
<td>M1a</td>
<td>spread to lymph nodes beyond regional ones</td>
</tr>
<tr>
<td>M1b</td>
<td>spread to bone</td>
</tr>
<tr>
<td>M1c</td>
<td>spread to other sites (regardless of bone involvement)</td>
</tr>
</tbody>
</table>

The TNM classification is an internationally recognised cancer staging system and the widespread use of screening means that most men are diagnosed with localised PCa.
However, some men are still diagnosed with more advanced disease, including some who present for the first time with metastatic disease.

The Gleason scoring system is the most widely used system for grading of PCa, and is based on histological pattern of tumour. It is the sum of two predominant patterns (grades 1-5; Table 2.2), which range from 2 to 10 (Gleason 1992). The higher the Gleason score, the more aggressive the cancer is.

**Table 2.2: Gleason patterns (Gleason 1992)**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Histological grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td></td>
<td>The cancerous prostate cells closely resemble normal prostate cells. The glands are small, well-formed, and closely packed.</td>
</tr>
<tr>
<td>Pattern 2</td>
<td>3 + 3</td>
<td>The glands are larger and have more tissue between them</td>
</tr>
<tr>
<td>Pattern 3</td>
<td>3+4 / 3+5</td>
<td>The tissue still has recognizable glands, but the cells are darker. Some cells have left the glands and have started to invade the surrounding tissue.</td>
</tr>
<tr>
<td>Pattern 4</td>
<td>4+3 / 5+3</td>
<td>The tissue has few recognizable glands. Many cells are invading the surrounding tissue</td>
</tr>
<tr>
<td>Pattern 5</td>
<td>4+4 / 4+5 / 5+4 / 5+5</td>
<td>The tissue does not have recognizable glands. There are often just sheets of cells throughout the surrounding tissue.</td>
</tr>
</tbody>
</table>

According to the D’Amico risk classification, the most commonly used risk stratification system, patients are grouped into low-risk, intermediate risk, and high-risk groups. Risk group classifications consider the PSA level, the Gleason score, and the clinical stage of the tumour. The low-risk patients have a PSA $\leq 10$ ng/ml, tumour stage $\leq cT2a$, and Gleason score $\leq 6$. Intermediate-risk patients have a PSA of 10–20 ng/ml, tumour stage is cT2b, and Gleason score of 7. High-risk patients have PSA $>20$ ng/ml and/or tumour stage $\geq cT2c$ and/or Gleason score of 8–10 (D'Amico, Whittington et al. 1998).

The prostate cancer prognosis classification as defined by AJCC 7th Edition (2009) involves the TNM classification along with the staging data, the Gleason Score and the PSA levels (AJCC 2009). We used this AJCC 7th Edition Classification for our study, wherein a
staging as ≥T2c, PSA score of ≥20ng/ml and Gleason Score of ≥8 were considered as Aggressive or high-risk PCa.

2.1.4 Risk factors of PCa

The exact etiology of PCa is not yet known, but apart from the influence of hormones (Nelles, Hu et al. 2011), the three major associated risk factors are age, race and familial history (Attard, Parker et al. 2016). Sadly, these factors cannot be altered after birth of an individual, and this makes the study for a “one-size-fits-all” biomarker for PCa a very challenging one.

Of the three, however, it is a well-established fact that PCa is mainly a disease of older men (Vaidyanathan, Karunasinghe et al. 2016). In the year 2013, it is estimated that around 64% of the men diagnosed with PCa were older than 65 years of age (Ministry of Health 2016). Family history of incidence of PCa also increases the risk of developing PCa (Vaidyanathan, Naidu et al. 2017). Having at least one first-degree relative with PCa doubles the risk of the disease, and interestingly, the risk is identified to further increase in case PCa was diagnosed in a family members at a relatively young age (≤65 years of age) and/ or if it affects more than three family members (Johns and Houlston 2003). Most importantly, however, with twice the risk observed in monozygotic twins than in dizygotic twins, and due to a higher incidence in African Americans compared to the Americans of Asian ancestry there is a strong indication of the role of genetics as an important parameter in the overall risk of PCa (Zeigler-Johnson, Spangler et al. 2008; Benafif and Eeles 2016). Further, it has also been demonstrated that among men with a germline mutation in the Breast Cancer 2 (BRCA2) gene (Castro and Eeles 2012; Attard, Parker et al. 2016) and Breast Cancer 1 (BRCA1) gene have a much higher risk of developing PCa (Castro and Eeles 2012). Apart
from the genetic analyses, migration studies have also provided evidence that environmental factors play a role in PCa etiology (Powell 2011).

There has been substantial amount of work done on hormonal predictors of PCa (Sofikerim, Eskicorapci et al. 2007; Key 2014). Androgens are not only necessary for the development but are also crucial for the functioning of a prostate gland (Sofikerim, Eskicorapci et al. 2007). However, the specific association of serum testosterone and pituitary hormone levels with the initiation and/or progression of PCa is not yet completely understood, and thus not looked into details in this thesis.
2.2. Prostate Cancer and lifestyle factors

Fast paced urbanization, and even changes in work-environments have had a dramatic impact on the lifestyle of individuals worldwide (Cordain, Eaton et al. 2005). Various lifestyle factors have already been demonstrated to contribute substantially to the development of cancers such as of the breast, prostate, and colorectal cancers (Lund Nilsen, Johnsen et al. 2000; Vaidyanathan, Naidu et al. 2017). Interestingly, along with influencing PCa risk, changes in lifestyle factors can influence the risk of PCa progression in men diagnosed with low grade PCa (Ornish, Weidner et al. 2005).

The major lifestyle factors studied and detailed by us across the thesis for their influence on the risk of PCa, especially aggressive PCa, are tobacco smoking, alcohol consumption, high BMI, and the influence of dietary factors consistent with New Zealand scenario, such as poor selenium in the diet, and high consumption of red meat.

2.2.1. Smoking status

Tobacco smoking is a major risk factor for many different types of cancers, promoting it as one of the most important modifiable and lifestyle-dependent risk factors for malignant diseases (Henriquez-Hernandez, Murias-Rosales et al. 2009; Ray, Henson et al. 2010). Having stated thus, it is also worthy of mention that the association between tobacco smoking and PCa remains controversial (Muller and Moreira 2011). Although many studies have investigated into the relationship between the status of lifestyle factors including tobacco smoking and PCa mortality, most studies have reported an overall increase of 30% in PCa mortality for current smokers as compared to non-smokers (Zu and Giovannucci 2009; Kenfield, Stampfer et al. 2011), and some other studies have also demonstrated a statistically significant positive association between smoking status and the risk of aggressive PCa (Spitz, Strom et al. 2000; Kobrinsky, Klug et al. 2003; Zu and Giovannucci 2009). More recently,
Weinmann et al. (2010) have also found a positive association between risk of PCa and current smokers (Weinmann, Shapiro et al. 2010). It is also worthy of mention that some studies have also found no correlation between smoking status and PCa risk (Ray, Henson et al. 2010).

This inconsistency in reported effect of tobacco smoking on risk of PCa, especially aggressive PCa arises from the fact that this is a self-reported parameter, and at times individuals also change their lifestyles and report false information about the history of tobacco smoking. It is also very difficult to classify and compare between different levels of cigarette smoking, as many unique factors such as quantity smoked, actively smoked or passively, the chemical composition used by the different brands, the kinds- whether rolled, or filtered, to name some, that need to be considered (Saha, Bhalla et al. 2007). The most common comparison, however, is between current smokers and recent quitters, and between non-smokers and past-smokers (with no history of smoking tobacco for more than 10 years) (Chen, Kuo et al. 2015) and was thus considered by us for this current study. However, studies have shown that non-smokers and past-smokers who had quit smoking for over 10 years may have a similar risk of PCa (Muller and Moreira 2011).

There are several possible mechanisms by which tobacco smoking can promote PCa carcinogenesis (Muller and Moreira 2011). These include, and are not restricted to, increased exposure to carcinogenic compounds present in tobacco smoke (Kidd, Hein et al. 2011). Smokers are often identified with increased levels of circulating androgens, such as testosterone and DHT, as compared to non-smokers (Dai, Gutai et al. 1988; Field, Colditz et al. 1994). Smokers also have higher levels of sex hormone-binding globulin (SHBG) which modulates androgen levels (Dai, Gutai et al. 1988; Field, Colditz et al. 1994; Roddam, Allen et al. 2008) and can affect the testosterone levels and affect various important gene pathways pertaining to PCa progression. Cigarette smokers are generally also observed to have lower
levels of serum selenium and insulin-like growth factor 1 (IGF-1) which are positively associated with PCa risk (Renehan, Zwahlen et al. 2004; Roddam, Allen et al. 2008).

2.2.2. Energy consumption and Body Mass Index

BMI is an estimate of an individual’s body fat based on their weight and height. It is calculated as a ratio of mass (in kg) to height (in m²). Based on the World Health Organization (WHO) classification, BMI is categorized as: <18.5 = underweight; 18.5-24.9 = normal weight; 25–29.9 = overweight; 30–34.9 = obese (class I); 35–39.9 = obese (class II); ≥40 = obese (class III) (Expert Panel on the Identification 1998). BMI is affected by the overall ratio of energy consumption (diet) to the energy expenditure (physical activity). A modern Western-style diet is characterized by high fat, cholesterol, and sugar intake due to the inclusion of various substances such as processed flours, sugars, meat and meat-products, and lack of fruits and vegetables. Adopting a Western-style diet may lead to an increased risk of obesity (high BMI) and has also been linked to increased PCa risk (Pollard and Luckert 1986; Llaverias, Danilo et al. 2010). Having said about Western-style diet, we would also like to mention that in Western countries, it is also generally observed that the people have a longer lifespan, but although life expectancy has increased in the recent past, health inequalities persist at a global scale (WHO 2016). Since Age is an important risk factor for PCa, this should not be forgotten. Obesity is associated with abnormal levels of sex hormones, inflammatory cytokines, and insulin-like growth factor (Burton, Tilling et al. 2010; Discacciati, Orsini et al. 2011).

High BMI (in men) has been demonstrated to have an effect on the circulating levels of androgen (Freedland and Platz 2007). Androgens are essential for the development and differentiation of normal prostate tissue, and its stimulation plays a crucial role in the initiation and progression of PCa (Platz and Giovannucci 2004). The blockade of androgen stimulation by lowering circulating androgen levels is a common treatment for advanced
metastatic PCa and those with high risk tumours in combination with radiation therapy (Miyamoto, Messing et al. 2004; Pagliarulo, Bracarda et al. 2012). Men with high BMI have an increased amount of adipose tissue, which produces aromatase, which aids in the conversion of testosterone to oestrogens, potentially leading the obese men to have lower testosterone levels (Freedland and Platz 2007). Despite the successes of lowering androgen level as a treatment for PCa, certain studies have found that lower serum testosterone concentrations were associated with an increased risk of more aggressive PCa (Platz, Leitzmann et al. 2005; Severi, Morris et al. 2006). It is possible that the decreased testosterone level promotes the selection of androgen-independent and more aggressive forms of PCa (Burton, Tilling et al. 2010; Discacciati, Orsini et al. 2011), but further work needs to be done in this regards.

It is generally observed that patients who are obese have a greater energy intake as compared to expenditure, leading to increased metabolic activity thus generating more reactive oxygen species (ROS) too. This can also lead to an increase in oxidative stress, resulting in the chronic low levels of inflammatory state observed among the obese individuals (Codoner-Franch, Valls-Belles et al. 2011). Inflammation has also been linked as a causal agent in various cancers (De Marzo, Platz et al. 2007; Birbach, Eisenbarth et al. 2011). Thus, men diagnosed with chronic prostatitis have an increased risk of developing PCa, while men taking non-steroidal anti-inflammatory drugs (NSAIDs) for a prolonged period have a decreased risk of PCa (De Marzo, Platz et al. 2007; Birbach, Eisenbarth et al. 2011). Circulating levels of inflammatory cytokines such as VEGF, IL-6 and TNF-a have been linked to PCa incidence (Burton, Tilling et al. 2010). Prostatic inflammation may promote PCa development and progression via the inflammatory cytokines (Radhakrishnan, Chachadi et al. 2011; Weiss, Simak et al. 2011).
Although there are several mechanisms linking high BMI with PCa as afore mentioned, direct association between BMI and risk of PCa still remains controversial. Putnam et al. (2000) and Engeland et al. (2003) have demonstrated statistically significant positive associations between high BMI and the risk of onset of PCa (Putnam, Cerhan et al. 2000; Engeland, Tretli et al. 2003); however, in contrast, Littman et al. (2007) and Discacciati et al. (2011) have discussed their findings pertaining to a statistically significant inverse association between high BMI and risk of PCa (Littman, White et al. 2007; Discacciati, Orsini et al. 2011); and there are certain other studies that have found no statistically significant association between high BMI and risk of PCa (Hubbard, Rohrmann et al. 2004). Kopp et al. demonstrated that obesity was positively associated with an increased prostate volume in men with localized PCa (Kopp, Han et al. 2011).

Certain studies (Giovannucci, Liu et al. 2007; Rodriguez, Freedland et al. 2007; De Nunzio, Freedland et al. 2011) have also indicated that obesity is statistically significantly associated with more aggressive PCa (classified by high Gleason score, tumour invasiveness, and/or patient mortality), while other studies have shown no associations (Wallstrom, Bjartell et al. 2009; Stocks, Hergens et al. 2010). However, these comparisons are further complicated as not all the subjects in these studies mentioned are of the same ethnicity and vary with respect to stratification by other risk factors such as tobacco smoking status.

2.2.3. Alcohol consumption

Over the past few decades there have been several studies (McGregor, Courneya et al. 2013; Demoury, Karakiewicz et al. 2016), reviews and meta–analyses (Perdana, Mochtar et al. 2016; Zhao, Stockwell et al. 2016) conducted to examine the association of risk of PCa with alcohol consumption, however, there still is a lack of proper evidence to support or negate the viewpoint. There are many limitations to any study pertaining to alcohol
consumption and risk of chronic diseases such as the kind of alcohol consumed, how much, and for how many years or months has an individual been consuming alcohol, to name some.

2.2.4. Dietary factors

Dietary factors have been linked with PCa risk and aggressiveness since quite some time now (Pollard and Luckert 1986; Llaverias, Danilo et al. 2010). PCa risk for example, of Asians living in North America has been observed to be greater than in Asia, this is likely due to adopting a high fat Western-style diet (Angwafo 1998). Certain dietary factors and micronutrients have also been shown to be associated with risk of PCa.

Selenium is a very important dietary supplementing mineral found in soil and when transformed to selenoproteins majority functions as an anti-oxidants (NRVs 2006). There are ~30 different types of selenoproteins found in the mammalian systems (NRVs 2006). Selenium is mainly absorbed from diet as selenomethionine and selenocysteine (Whanger, Vendeland et al. 1996). The recommended selenium daily intake for adult men in New Zealand and Australia is 70ug/day, with the upper level of intake at 200ug/day (NRVs 2006). In No Observed Adverse Effect Level (NOAEL) Study, selenium toxicity was not observed at 800ug/day (Longnecker, Taylor et al. 1991).

The level of selenium present in food is dependent on soil selenium levels which can vary widely across different geographical regions (Tinggi 2008). Most of New Zealand has soil levels of selenium at <0.6 mg/kg (Figure 2.1) which is considered as selenium deficient (Karunasinghe, Ferguson et al. 2006). In contrast, very high levels of selenium intake can increase the risk of aggressive PCa (Chan, Oh et al. 2009; Steinbrenner, Speckmann et al. 2011).

The potential chemo-preventative properties of selenium on PCa was first demonstrated by Clark et al. (1996), who found a significant reduction in the incidence of PCa and even reduced mortality when participants were supplemented with selenium for an average of
slightly more than 4 years (Clark, Combs et al. 1996). Certain other studies have also found selenium supplementation to have a chemo-preventative impact on PCa (Duffield-Lillico, Dalkin et al. 2003; Kim, Sun et al. 2005; Sabichi, Lee et al. 2006; Pourmand, Salem et al. 2008). Other studies, such as the Selenium and Vitamin E Cancer Prevention Trial (SELECT), one of the largest randomized clinical trials with 35,000 subjects, has demonstrated that selenium supplementations and/or serum selenium levels are not significantly associated with the risk of PCa (Lippman, Klein et al. 2009; Dennert, Zwahlen et al. 2011; Marshall, Tangen et al. 2011). This study, however, used a different form of selenium supplementation and a study population with a higher level of selenium, thus making any comparison futile.

Certain factors such as cigarette smoking can negatively impact selenium’s chemo-preventative effect. Cigarette smoking can produce Cadmium, which forms a complex with Selenium, thereby removing its anti-oxidant effect (Schopfer, Drasch et al. 2010).

The potential chemo-preventative properties of selenium on PCa was first demonstrated by Clark et al. (1996), who found a significant reduction in the incidence of PCa and even reduced mortality when participants were supplemented with selenium for an average of slightly more than 4 years (Clark, Combs et al. 1996). Certain other studies have also found selenium supplementation to have a chemo-preventative impact on PCa (Duffield-Lillico, Dalkin et al. 2003; Kim, Sun et al. 2005; Sabichi, Lee et al. 2006; Pourmand, Salem et al. 2008). Other studies, such as the Selenium and Vitamin E Cancer Prevention Trial (SELECT), one of the largest randomized clinical trials with 35,000 subjects, has demonstrated that selenium supplementations and/or serum selenium levels are not significantly associated with the risk of PCa (Lippman, Klein et al. 2009; Dennert, Zwahlen et al. 2011; Marshall, Tangen et al. 2011). This study, however, used a different form of
selenium supplementation and a study population with a higher level of selenium, thus making any comparison futile.

Certain factors such as cigarette smoking can negatively impact selenium’s chemopreventative effect. Cigarette smoking can produce Cadmium, which forms a complex with Selenium, thereby removing its anti-oxidant effect (Schopfer, Drasch et al. 2010).
2.3. Prostate Cancer and Genetic factors

Familial history, an important risk factor for PCa, suggests a strong genetic basis for the disease (Matikaine, Pukkala et al. 2001; Stanford and Ostrander 2001). In an analysis comparing the PCa development between monozygotic and dizygotic twins, the monozygotic twins were identified to have a significantly higher concordance rate in the overall risk of developing PCa compared to the dizygotic twins (Gronberg, Damber et al. 1994; Page, Braun et al. 1997; Lichtenstein, Holm et al. 2000). This suggestion about the possibility of existence of PCa susceptibility genes has promoted the role of genetic and molecular diagnostics in the field of biomedical research for the diagnosis of PCa. Knowledge of the role of genetics can help with the understanding of pathology, identification, and development of treatment regimens for PCa. Genetic involvement in PCa including the role of somatic copy number alterations, structural rearrangements, point mutations, SNPs, long non-coding RNAs, and micro-RNAs have been reviewed by a number of research groups (Batra, Girdhani et al. 2014; Wallis and Nam 2015). Predicting an individual’s risk of developing PCa may also motivate high-risk men to increase screening, and/or select the suitable less-invasive treatment (Lindor, McMaster et al. 2008).

2.3.1. Genome Wide Association Studies

GWAS examine genetic variants in populations with similar traits to see if any genetic variants are associated with the investigated traits. There has been a considerable amount of published literature about GWAS-related research carried out by various groups pertaining to risk of PCa (Kote-Jarai, Easton et al. 2008; Kote-Jarai, Amin Al Olama et al. 2011; Kote-Jarai, Olama et al. 2011; Tao, Wang et al. 2012; Amin Al Olama, Kote-Jarai et al. 2013; Kote-Jarai, Saunders et al. 2013). The current focus is to identify the interactions between PCa associated SNPs and known environmental risk factors in modulating PCa
aggressiveness, as well as other SNPs effect on gene expression (Lindstrom, Schumacher et al. 2011; Chen, Ren et al. 2013).

2.3.2. Single Nucleotide Polymorphism Genotyping

SNPs are the DNA sequence variations that occur when a single nucleotide in the genome differs within a population. Genetic variations are a major factor contributing to the differences in disease susceptibility amongst individuals (Tweeddy and Belmont 2009). SNPs are the most common genetic variations in the human genome and contribute to a variety of phenotypes (Anderson, Hansen et al. 2006). There has been a series of publications that have been used to demonstrate associations between disease susceptibility and SNPs (Orr and Chanock 2008; Ciampa, Yeager et al. 2011; Liu, Wang et al. 2011; Pomerantz, Werner et al. 2011). Importance of both synonymous and non-synonymous SNPs in disease etiology and their mechanisms of interaction have also been broadly reviewed (Hunt, Sauna et al. 2009; Chen, Davydov et al. 2010; Zhao, Han et al. 2014). Similarly the association of both, exonic as well as intronic SNPs in PCa and its various states have also been well reported (Hazelett, Rhie et al. 2014; Jin, Jung et al. 2016; Whittington, Gao et al. 2016). Effect of SNPs in the variation of transcription factor (Jin, Jung et al. 2016; Kumar, Ambrosini et al. 2017) and miRNA binding and the subsequent effects on cancer etiology including that of PCa have also been well researched (Lose, Srinivasan et al. 2012; Preskill and Weidhaas 2013).

GWAS help in identifying large number of SNPs across a number of populations which are associated with a risk of PCa (Witte 2009). The statistical significance demonstrates the effects of SNPs on risk of PCa, and thereby promoting SNP genotyping as a potential tool to identify the susceptible polymorphism. SNP genotyping can also be used to define the risk of progression of PCa in individuals (Lin, FitzGerald et al. 2011). The identification of SNP biomarkers can potentially help identify the potentially unsuspected pathogenic mechanisms, thereby helping develop better ways of detecting PCa (Xu, Wise et al. 2010).
There are several different technologies that can be employed for SNP genotyping aiding various large-scale genotyping studies. SNP genotyping studies help with GWAS in understanding the data in various case-control studies better. SNP genotyping utilizes a very straight-forward approach.

2.3.3. Methods for detecting SNPs

The SNP genotyping methods can predominantly be classified based on the chemicals and the properties of the chemicals used in the respective protocols. The two broad classifications are the differential-hybridization method, which utilizes a pair of specific oligonucleotide probes which are designed specifically so that the alleles located at the SNP can bind to either one of the two probes (Kwok, 2001; Syvanen, 2001); and the enzyme-based methods, which involves endonuclease enzymes that can recognize and cleave specific DNA sequences (Ding and Jin 2009). Certain examples of the differential hybridization method includes Affymetrix array and Dynamic allele-specific hybridization (DASH) genotyping (Ding and Jin 2009; Bichenkova, Lang et al. 2011), and Restriction fragment length polymorphism (RFLP) is one of the earliest methods used for the identification of SNPs, which uses an enzyme-based method (Todd, Donoff et al. 2001). TaqMan assays take advantage of both differential and enzyme-based hybridization (Shen, Abdullah et al. 2009). It makes the genetic make-up product- SNP, where the nucleotide extension (using a visualization method that is large, depending on the selection used on the adjustment platform), can be expressed using a mass spectrometer (Shen, Abdullah et al. 2009). The selection of the most appropriate SNP genotyping method is dependent on the number of samples and the SNPs to be screened.

We used Sequenom MassARRAY iplex, a form of differential-hybridization method, to genotype the SNPs of interest. Not only is it a highly sensitive technique, but is also specifically designed to rapidly distinguish between various genotypes by combining the iPLEX gold chemistry along with matrix-assisted laser desorption ionization-time of flight
mass spectrometry (MALDI-TOF MS). The MassEXTEND primer extension chemistries combined with high-density SpectroCHIP arrays allows high-throughput analysis of upto 40 SNPs in just a single reaction (Wright, Heggarty et al. 2008; Gabriel, Ziaugra et al. 2009). iPLEX MassARRAY is also a cost-effective genotyping tool with very low error rate. The cluster plots generated by iPLEX MassARRAY were then manually checked for the quality. Certain random examples are provided in the Appendix section of the thesis.

2.3.4. Statistical analysis

One of the most important steps in SNP genotyping is to analyse the data and interpret it. The raw data that is generated by any chip contains identifying information for both- the individual and the marker (SNP ID) which corresponds to the genotype. The output files should then be analyzed to assess the levels of association between the phenotype and the genetic variation(s) amongst SNP genotypes across all the individuals (or, samples) that were typed. There are several methods, programs and even mathematical models to ascertain association of SNP genotypes to samples based on the requirements of the researcher(s) and/or study design. In depth details of the same is beyond the scope of this thesis, and therefore not detailed here. The basic protocol, however, is outlined below.

To start with, certain quality control steps should be undertaken to ensure that the raw data being assessed for association with the disease of interest (PCa, in our case) is robust and bereft of any bias. It is a good practice to confirm the genotyping call rate is accurate and consistent. The allele frequencies also need to be checked to see if the sample is in compliance with Hardy-Weinberg equilibrium (HWE) and is not in linkage (Burton et al., 2007; McCarthy et al., 2008). Since Mendelian Randomization and haplotype analyses are not feasible for our dataset, as the various polymorphisms, as per Smith and Ebrahim (2003),
may have several phenotypic effects associated with disease(s), these were not considered in this analysis (Smith and Ebrahim 2003).

2.3.4.1. PLINK based Identity-by-state (IBS) clustering

PLINK is a very useful whole-genome association and linkage analysis toolset that was developed by Purcell et al., (2007) around ten years back, but is still one of the most reputable bioinformatics analytical tools, especially while handling for GWAS data analysis to date (Purcell, Neale et al. 2007). One of the important benefits of using this software is that it provides a wide range of tools, including GWAS data manipulation, quality control (HWE and linkage disequilibrium), association and interaction studies, GWAS-based meta-analysis, epistasis (Purcell, Neale et al. 2007) to name some.

2.3.4.2. Artificial Neural Networks

Artificial Neural Networks (ANNs) or the connectionist system is a computational model, based on a large section of individual artificial neural units, which can be considered to be similar to the observed behavioral pattern of a biological brain’s axons (Hall, Frank et al. 2009). Just like in the biological brains, each neural unit in ANNs is also inter-connected with many other neural units, and these links can not only enhance but also inhibit the activation state of the adjoining neural units (Hall, Frank et al. 2009). Each individual neural unit and its function and/or interaction can be computed using the summation function. These systems are recognized to be “self-learning” and “self-trained”, and not explicitly programmed (Witten, Frank et al. 2016). The systems excel in areas of study where the solution and/or feature detection is difficult to express in any other “classical” computer program. ANNs is typically of a cube design consists of multiple layers, and the signal path is observed to traverse from the first (or, input), to the last (or, output) layer of the neural units (Hall, Frank et al. 2009).
One of the important tools employed to generate ANNs is WEKA (Hall, Frank et al. 2009). This workbench is a collection of machine-learning algorithms and aids in the data pre-processing. WEKA provides the user with the implementation(s) of learning algorithms that can easily be applied to various datasets. WEKA also includes several tools that is used for the transformation of datasets.

For the main data mining problems, WEKA can be employed for regression, classification and clustering, association of rule mining, and attribute selection (Hall, Frank et al. 2009). All the algorithms take their input in the form of a single relational table which can be read from a file or generated by a database query.

The ANN classifier within the WEKA- multilayer perceptron (MLP), employs a three-layer feedforward neural network accompanied by a sigmoid function (Hall, Frank et al. 2009). MLP is a non-linear neural network and is comprised of neurons that are classified in layers. This sigmoid function is employed to encode every net as an eight-bit binary unit. From recently, ANNs have been employed for diagnosing and predicting potential cancer biomarkers. A feedforward ANN with back propagation learning model is implemented with input, hidden and output layer.
2.4. Chapter Summary

PCa is a crucial health concern worldwide including New Zealand. There are considerable numbers of known symptoms of this disease, but not all cases are detected early on. The PSA level, Gleason score, and staging information are the prevalent criterion for diagnosing and categorizing this disease. However, with PSA not being completely error-free, various molecular diagnostics is the next emerging method of diagnosing PCa.

To better understand the disease incidence and its prognosis, it is undoubtedly important to focus on the important risk factors such as age, race and familial history. However, we believe that these risk factors have an underlying flavor of lifestyle factors, and were thus discussed herewith. Knowledge of various Gene x Environment interactions is crucial in understanding the prognosis of any disease, and thus demands acute attention when planning a case-control study.

Bearing these in mind, a well-planned genetic study is required to better understand the local conditions and the dependent genetic factors followed by a thorough analysis to identify a biomarker for this critical disease, and was thus the basis of this thesis.
CHAPTER THREE

Hypothesis, Aims and Plot of the Thesis

“The number of natural hypothesis that can explain any given phenomena is infinite.”

- Albert Einstein
Chapter 3. Hypothesis, Aims and Plot of the Thesis

3.1. Rationale of the Study

Prostate cancer (PCa) is highly prevalent, and around 1 in 6 patients are at risk of developing the aggressive form of the disease (Cooperberg, Vickers et al. 2010). It has become one of the most significant male health concerns worldwide (Jemal, Bray et al. 2011). The patients with high PSA are generally asked to undergo biopsy, which could lead to post biopsy complications even in people diagnosed negative for PCa. It is crucial that we exploit the available resources such that we can identify a reliable biomarker for PCa so that every individual is not made to undergo the same treatment regime.

The risk factors for PCa include association of demographic and lifestyle factors, age, and genetics. Owing to the fact that PCa is such a multifactorial disease and gene x environment prospects play such a crucial role in the varied expression of genes in PCa (Vaidyanathan, Naidu et al. 2017), we hypothesized that multi-dimensional approach of SNP genotyping in participating patients vs control and then analysing the results by using combinatorial bioinformatics methods such as P-link and ANN by adjusting to various environmental effects would lead us to unearthing the much needed biomarkers for PCa especially, aggressive PCa. We have herewith considered the association of demographic and lifestyle factors, age, and SNP x SNP interactions as a risk of aggressive PCa.

3.2. Limitations of the Study

Just like other researchers working on SNP association with risk of PCa (Henriquez-Hernandez, Murias-Rosales et al. 2009), we are conscious of the fact that the power analysis of the cohort size for a study of our scale was not performed at the start of the study. However, as per Hong and Park (2012), a lower sample size can be used for studying common diseases (PCa, in New Zealand, in this case) that will still give desirable significant
statistical power. Also, statistical power increases by having a higher number of controls per cases (as in case of our study) (Hong and Park 2012).

We used the software Statistical Package for the Social Sciences (SPSS) to analyse the power of our data ad hoc for age and BMI- the only two continuous variables in our study, and found that our sample size has statistical power for a case-control study (Table 3.1 for the age of the sample size and Table 3.2 for the BMI of the sample size).

**Table 3.1: Power calculation for age as a parameter for prostate cancer vs healthy controls**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Value</th>
<th>F</th>
<th>Hypothesis df</th>
<th>Error df</th>
<th>Sig.</th>
<th>Noncent. Parameter</th>
<th>Observed Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAA_POWERR_Age</td>
<td>.299</td>
<td>108.014&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.000</td>
<td>253.00</td>
<td>.000</td>
<td>108.014</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.701</td>
<td>108.014&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.000</td>
<td>253.00</td>
<td>.000</td>
<td>108.014</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.427</td>
<td>108.014&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.000</td>
<td>253.00</td>
<td>.000</td>
<td>108.014</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>.427</td>
<td>108.014&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.000</td>
<td>253.00</td>
<td>.000</td>
<td>108.014</td>
<td>1.000</td>
</tr>
</tbody>
</table>

a. Design: Intercept
Within Subjects Design: GWAA_Power_Age
b. Exact statistic
c. Computed using alpha = .05
Table 3.2: Power calculation for BMI as a parameter for prostate cancer vs healthy controls

<table>
<thead>
<tr>
<th>Effect</th>
<th>Value</th>
<th>F</th>
<th>Hypothesis df</th>
<th>Error df</th>
<th>Sig.</th>
<th>Noncent. Parameter</th>
<th>Observed Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAA_Power_BMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>.047</td>
<td>12.574b</td>
<td>1.000</td>
<td>253.000</td>
<td>.000</td>
<td>12.574</td>
<td>.942</td>
</tr>
<tr>
<td>Wilks' Lambda</td>
<td>.953</td>
<td>12.574b</td>
<td>1.000</td>
<td>253.000</td>
<td>.000</td>
<td>12.574</td>
<td>.942</td>
</tr>
<tr>
<td>Hotelling's Trace</td>
<td>.050</td>
<td>12.574b</td>
<td>1.000</td>
<td>253.000</td>
<td>.000</td>
<td>12.574</td>
<td>.942</td>
</tr>
<tr>
<td>Roy's Largest Root</td>
<td>.050</td>
<td>12.574b</td>
<td>1.000</td>
<td>253.000</td>
<td>.000</td>
<td>12.574</td>
<td>.942</td>
</tr>
</tbody>
</table>

a. Design: Intercept
Within Subjects Design: GWAA_POWER_BMI
b. Exact statistic
c. Computed using alpha = .05

Thus, since the power of our statistical test is sufficient (i.e., is .80 or higher), the sample size is sufficient for the research purposes discussed here (Thompson 2006).

We are also aware of the fact that a number of SNPs across many more genes can and should be looked into to identify potential SNP biomarkers for PCa. We could not add a number of genes, even if they were suggested in literature. This is because multiple primers, based on similar molecular weights, for SNP genotype were pooled to be used in a multiplexing process, and run on the machine. Due to limited resources, further genotyping using single-run methods such as Taqman Assay was not pursued for the failed SNPs. It was difficult to get the right combination of SNPs that could all be compatible. It is worthy of mention that we involved additional SNP genotype data from the Urological database (ACSR, University of Auckland) also to this study. Our literature search and subsequent SNP genotype data generation was completed by the year 2013, so, we have not included some SNPs from the recent GWAS publications in this research.
We also acknowledge the fact that ours being a Caucasian cohort based SNP analysis, one of the most crucial limitations is the reproducibility of our analyses in other populations within or outside New Zealand among patients with the same disease, or other similar diseases. However, we would like to suggest that this be used as a preliminary study in designing similar cohorts from New Zealand or elsewhere.

Although the technologies for SNP genotyping and subsequent data analysis have been improved rapidly in the last few years, merging new technologies is challenging for researchers (Chen and Sullivan 2003). Moreover, with the advent of revolutionizing genomic research tools such as next generation sequencing that can sequence an entire human genome within just a single day (Behjati and Tarpey 2013), SNPs genotype techniques may soon be out of favour. This does seem to be a major limitation of this thesis, but, we believe that SNP genotype and the subsequent data analyses, as explained in the chapters of this thesis, is a good start-point for further work leading up to identification of a SNP biomarker for aggressive PCa.
3.3. Aims of the thesis

- To assess the risk of aggressive prostate cancer in a New Zealand cohort using established SNPs for cancer and/or prostate cancer.
- To assess the genome environment interactions in causing risk of aggressive prostate cancer.
- To understand machine-learning tools for assessing SNP genotype data.
- To compare the aggressive prostate cancer risk analysis by machine-learning, and conventional statistical protocols to narrow down the list of potential SNP biomarkers for the current study.
- To validate the effect of the risk genes using protein quantitation in cell-lines.
3.4. Thesis Layout

This thesis contains ten chapters and appendices summarized as follows (Figure 3.1):

**Chapter One- Introduction:** This chapter introduces the topic and plot to the readers.

**Chapter Two– Review of Literature:** This chapter provides a literature review addressing the gravity of the disease of interest- PCa. It summarizes the present-day scenario, and the direction in which the current research will proceed.

**Chapter Three– Hypothesis, Aims, and Plot of the thesis:** This chapter justifies the rationale behind this study with an outline of the research objectives to conduct this research.

**Chapter Four– Materials and Methods:** The general materials and methodologies used to perform this study are written in details in this chapter.

**Chapter Five– SNP Genotyping:** This chapter is divided into four segments, based on the various publications begotten from our research. The gene x environment (G X E) scenario, role of ageing in risk of PCa, the (G X E) x age scenario, and the SNP-SNP interaction and the risk of aggressive PCa have been discussed here.

**Chapter Six– Machine-Learning Analysis:** The results obtained from machine-learning analysis of the SNP Genotyping data is explained in this chapter written in the style of a research article.

**Chapter Seven– Validation of SNP Genotyping results:** The results obtained by analysing the protein-level expression of the genes identified in Chapters 6 are explained in this chapter written in the style of a research article.

**Chapter Eight– Final Discussion:** This chapter presents the final discussion of the thesis.

**Chapter Nine– Future Directions:** This chapter is used to explain the importance of various individuals involved with the assessment of PCa risk for better patient outcome, and the way to look forward in combatting this serious illness.
Chapter Ten - Bibliography: This chapter comprises of all the references cited across various publications and chapters in this thesis.

Appendices: The list of SNPs studied, various map-plates of the results obtained, screenshots of the HWE and LD analyses and Co-authorship forms are also provided in the appendices.

![Figure 3.1 PhD Thesis Flow Chart](image-url)
CHAPTER FOUR

Materials and Methods

“The real purpose of the scientific method is to make sure Nature hasn’t mislead you into thinking you know something you don’t actually know.”

-Robert M Pirsig
Chapter 4. Materials and Methods

The overall methods used in this research are SNP genotyping, data analysis using Identical-By-State (IBS) Clustering using PLINK and by Machine-learning, and analysis of the expression of proteins associated with important genes of interest in imported cell lines.
4.1. Patient recruitment for the study

Patients described in this study are those that took part in the ‘Genomic Studies on Prostate Cancer’ carried out at the University of Auckland, New Zealand with collaboration of the Urology Department, Auckland Hospital. Patients with clinically established diagnosis of PCa (both, aggressive and non-aggressive) from the Auckland, Regional Urology Registries (Auckland, Middlemore, and North Shore hospitals), and certain private practices in the Waikato region of New Zealand were invited to take part in this study. Recruitment was restricted to those above 40 years regardless of ethnicity and was carried out with informed consent between the years 2006 and 2014. The current analysis was restricted to patients with a European ancestry. Eventually, 254 PCa patients were considered for analyses. (Ethics reference NTY05/06/037 by Northern B Ethics Committee, New Zealand, previously Northern Y Ethics Committee, New Zealand).

Additionally, 369 males of European ancestry from the Auckland region, New Zealand, with no reported symptoms and/or clinical diagnosis of PCa or any other cancer except for skin cancers were considered as healthy controls for this study (Ethics reference NTY/06/07/AM04 by Northern B Ethics Committee, New Zealand, previously Northern Y Ethics Committee, New Zealand). These men are part of the participants who have previously participated in ‘Optimising Selenium for health benefits’ study conducted by the Auckland Cancer Society Research Centre, University of Auckland. These men have given consent to their blood/DNA samples and other data collected for the above study be used in other studies approved by a New Zealand accredited ethics committee. Only men in the age range 40-90y were considered in the current analysis.

Each individual participating in both studies completed a demographic and lifestyle questionnaire too.
4.2. Collection of human samples:

The patients’ blood samples were collected by registered nurses or trained phlebotomists at the respective outpatient clinics at various locations- Auckland, North Shore and Counties Manukau District Health Boards. Blood sample collection of Waikato patients were carried out at the Urology Waikato Clinic at Knox Street in Hamilton, New Zealand by trained staff. Blood was stored in EDTA BD Vacutainer (Becton and Dickinson Ltd. USA). The blood samples of healthy controls were collected by a trained phlebotomist appointed at the Faculty of Medical and Health Sciences, the University of Auckland, New Zealand or at the New Zealand Blood Bank, Great South Road Centre, Epsom, Auckland New Zealand.
4.3. SNP Selection:

To select SNPs that are associated with PCa risk, and especially aggressive PCa, a thorough literature search was performed using the NCBI database. The keywords used were “prostate cancer” and/or “aggressive prostate cancer”, along with “SNP genotyping sequenom massarray iplex”, and/or “genome-wide association studies (and/or GWAS)”. Since our cohort was of European ethnicity, GWAS targeting PCa-related genes in a primarily Caucasian population was considered. SNPs that were found to have a significant (p≤0.05) association with PCa (and/or, aggressive PCa) risk only were included for this study (Supplementary Table 1). Additionally, SNP data available at the Urology Database, Auckland Cancer Society Research Centre, Faculty of Medical and Health Sciences, University of Auckland were also used in the current analysis.

The flanking sequences of the SNPs of interest were obtained using the “Human Genome Browser Gateway” at [http://genome.ucsc.edu/cgi-bin/hgGateway](http://genome.ucsc.edu/cgi-bin/hgGateway) for Homo sapiens. Sequences of 200 base-pairs, both, upstream and downstream of the SNPs were retrieved. Common SNPs were replaced with “N” and the polymorphism was changed to “[_/_]”, where, “_” represents the 2 possible alleles of the same SNP. The flanking sequences were then loaded onto the MassARRAY® Assay Designer 3.1 with single base extension mode and a multiplex level between 24 and 40 (SEQUENOM 2006).

MassARRAY® Assay Designer 3.1 was employed to design the appropriate set of amplification primers and the set of extension primers for the SNPs. Care was taken to avoid primers that may form dimers during a multiplexed reaction. The melting temperature (T_m) was designed based on the number of bases considered, and the G-C content was also checked, as the triple bond between the bases Guanine (G) and Cytosine (C) increase the melting temperature, thereby affecting the whole PCR process (Mammedov, Pienaar et al. 2008).
The Assay Designer was also utilized to fit in as many SNPs into a single iPLEX batch as possible. For SNPs that could not be fitted into the same iPLEX batch, certain SNPs 10kb upstream and downstream, in strong linkage disequilibrium with the original SNPs, were used instead. The HapMap database (HapMap Genome Browser release #28; Phases 1, 2 & 3 - merged genotypes & frequencies) were used to retrieve information about the SNPs that were within 10kb upstream and downstream region of the SNP (originally planned). The genotypes retrieved were from the Caucasian European population Utah residents with Northern and Western European ancestry (CEU). The PLINK software version 1.07 (Purcell, Neale et al. 2007; Vaidyanathan, Naidu et al. 2017) was used to identify the SNPs that were in linkage equilibrium. The same software was later on used to identify for the SNPs that failed the Hardy Weinberg Equilibrium (HWE) after genotyping too (Vaidyanathan, Naidu et al. 2017). This is explained in detail in section 4.5.2 of this thesis chapter.
4.4. SNP Genotyping using iPLEX MassARRAY technology

4.4.1. Principle

4.4.1.1. DNA extraction using QIAcube™

DNA was extracted from the whole blood samples acquired from the patients and healthy controls who participated in this study. After the machine-run procedure, the DNA was first bound to the silica membrane, then the contaminants were removed with deferasirox (wash reagent), and the purified DNA was eluted from the silica membrane with the elution reagent provided with the kit.

The extracted DNA samples were tested for purity and concentration with a NanoDrop® ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA). The DNA templates are needed to be highly pure as the PCR process could be affected by the presence of RNA and/or protein contaminants. The ratio of absorbance reading, of the DNA extracted from the blood samples of the participants, between 260nm and 280nm wavelength was observed to be between 1.7 and 2.0 and the ratio of absorbance reading at 260nm and 230nm wavelength to be between 2.0 and 2.2.

4.4.1.2. iPLEX MassARRAY

The MassARRAY iPLEX Gold (Sequenom®) protocol employs a single base extension of 2',3' dideoxynucleotides (ddNTPs) at the allele of interest (SEQUENOM 2006). The two possible alleles in any individual SNP can be distinguished by the difference in the mass of ddNTPs using a mass spectrometry.

Since different SNPs that are studied have extension primers of varying lengths and masses, this allows differentiation of different SNPs in the same MassARRAY assay (SEQUENOM 2006). Each MassARRAY experiment can be used to analyse up to 40 SNPs at the same time (SEQUENOM 2006).
Detection of mass-specific products is achieved by using MALDI-TOF MS (Singhal, Kumar et al. 2015). For this project, the mass spectrometry was carried out at the Sequenom Facility, Liggins Institute, Faculty of Medical and Health Sciences, University of Auckland, Auckland. The mass spectrometry data can be translated into the genotype “calls” with known ddNTPs and extension primer mass via MassARRAY® Typer 4.0.2 Software (SEQUENOM 2006). Few examples of the genotype “calls” are provided in the Appendix.

The first step of an iPLEX MassARRAY reaction is PCR amplification of the sequence including and around the SNPs of interest. This makes multiple copies of the region harboring the SNP. The forward and reverse primers bind to the DNA sequences ~200bps upstream and downstream of the SNP of interest. HotStar Taq DNA Polymerase (for Seqeunom®) is activated at 95°C for 15 minutes and elongated forward and the reverse primers by annealing deoxynucleotide triphosphates (dNTPs) to the 3’ end of primers. Magnesium from MgCl₂ acts as a co-factor for DNA polymerase. The DNA polymerases were used to elongate and amplify the selected DNA sequence.

Once the PCR is completed, the remaining dNTPs are needed to be removed to avoid contamination and/or non-sense binding. This was done by using Shrimp Alkaline Phosphatase (SAP). SAP enzymes dephosphorylate dNTPs to deoxynucleotide diphosphates (dNDPs) rendering them unavailable for the following iPLEX extension reaction.

Then, the iPLEX extension reaction was carried out. The extension primers anneal to the amplified PCR products. The extension primers extend by one nucleotide with mass-modified ddNTPs complementary to the SNP sites.

Following this, resin desalting was carried out for optimization of mass spectrometric analysis. Sodium (22Da) and potassium (38Da) ions have a similar mass to A/C (24Da) and C/G (40Da) SNPs respectively and thus can be difficult to discriminate if present in the
solute. Resin forms a complex with the free ions thus retaining and removing them before the next step—Mass Spectrometry analysis.

### 4.4.2. Reagents and Equipments

The reagents used for SNP genotyping are mentioned in Table 4.1.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNAse/DNase free water</td>
<td>Gibco, Invitrogen Corporation, Grand Island, NY, USA</td>
</tr>
<tr>
<td>PCR Buffer</td>
<td>Sequenom, Herston, Australia</td>
</tr>
<tr>
<td>MgCl₂</td>
<td>QIAGEN, Victoria, Australia</td>
</tr>
<tr>
<td>Hotstar Taq</td>
<td>Applied Biosystems, Auckland, New Zealand</td>
</tr>
<tr>
<td>SAP</td>
<td>Sequenom, Herston, Australia</td>
</tr>
<tr>
<td>SAP buffer</td>
<td>Sequenom, Herston, Australia</td>
</tr>
<tr>
<td>iPLEX buffer</td>
<td>SIGMA-ALDRICH, Castle Hill, Australia</td>
</tr>
<tr>
<td>iPLEX termination mix</td>
<td>SIGMA-ALDRICH, Castle Hill, Australia</td>
</tr>
<tr>
<td>Ext primer mix</td>
<td>SIGMA-ALDRICH, Castle Hill, Australia</td>
</tr>
<tr>
<td>iPLEX enzyme</td>
<td>SIGMA-ALDRICH, Castle Hill, Australia</td>
</tr>
<tr>
<td>1kb ladder</td>
<td>Invitrogen, Victoria, Australia</td>
</tr>
<tr>
<td>2×loading buffer</td>
<td>SIGMA-ALDRICH, Castle Hill, Australia</td>
</tr>
</tbody>
</table>

### 4.4.3. Protocol

#### 4.4.3.1. DNA Extraction from EDTA blood

The QIAcube™ DNA extraction procedure employing the QIAamp® DNA Mini kit (Qiagen, USA) was run based on the instructions in the handbook. Each DNA sample was extracted from 300µl of blood serum. Preparation and clean-up were performed following the protocols provided in the QIAcube™ DNA Handbook.

#### 4.4.3.2. iPLEX preparation

All the laboratory consumables such as the 384-well plates, pipette tips, falcon tubes, etc., were exposed to UV-sterilization with a CL-1000 UV crosslinker (Ultra-Violet Products Limited) for ~20 minutes before use each time. A Master Mix was prepared in a Hoefer®
PCR Workstation. The workstation was also exposed to UV-sterilization for ~20 minutes before the preparation of the PCR Master Mix. The bench surface was cleaned with 70% ethanol before and after aliquoting in order to minimize contamination from DNA amplicons.

4.4.3.3. **DNA plating**

The extracted DNA samples were diluted to 10ng/µl in the 96-well plates. The diluted DNA aliquots (2µl per well), as per the requirement of the protocol (SEQUENOM 2006) were transferred to 384-well plates using a multichannel pipette. The plates were then centrifuged (Eppendorf Centrifuge 5810R, Global Science) at 1000rpm for 1 minute to ensure that the DNA aliquots were at the bottom of the wells without any air bubbles. The DNA aliquots were then completely dried at 55°C for 30 minutes using a thermal cycler (Veriti 384-well Thermal Cycler, Applied Biosystems). Our 384-well plates had negative controls (water), positive controls (HAPMAP Samples), and repeats scattered, so that the successful run of the plate could be verified.

4.4.3.4. **PCR reaction**

The PCR MM was prepared by following the recipe stated in Table 4.2 and then placed on ice. With a multichannel pipette, 5µl of the MM was aliquoted into each well of the 384-well plate. The plates were then sealed using a Polypropylene film (MicroAmp™ Clear Adhesive Film, Applied Biosystems) and centrifuged at 1000rpm for 1 minute at room temperature to remove any air bubbles. The plates were then checked and ensured that the volume in each well was the same. Following this, the sequenom PCR program was run in a Thermal Cycler (Veriti 384-Well Thermal Cycler, Applied Biosystems) (Table 4.3).
### Table 4.2 PCR Master Mix

<table>
<thead>
<tr>
<th>Reagents</th>
<th>x1 reaction (in µl)</th>
<th>384 plate (with some room for pipetting errors) (in µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>3.25</td>
<td>1300.00</td>
</tr>
<tr>
<td>PCR buffer 10x</td>
<td>0.625</td>
<td>250.00</td>
</tr>
<tr>
<td>MgCl 25mM</td>
<td>0.325</td>
<td>130.00</td>
</tr>
<tr>
<td>DNTPs (25mM)</td>
<td>0.10</td>
<td>40.00</td>
</tr>
<tr>
<td>HotStar Taq</td>
<td>0.20</td>
<td>80.00</td>
</tr>
<tr>
<td>Forward + Reverse Primers (1pmol/µl)</td>
<td>0.50</td>
<td>200.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5.00</strong></td>
<td><strong>2000.00</strong></td>
</tr>
</tbody>
</table>

### Table 4.3 PCR Programme

<table>
<thead>
<tr>
<th>PCR programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 95°C for 15 min</td>
</tr>
<tr>
<td>2: 94°C for 20 sec</td>
</tr>
<tr>
<td>3: 56°C for 30 sec</td>
</tr>
<tr>
<td>4: 72°C for 1 min</td>
</tr>
<tr>
<td>5: 72°C for 3 min</td>
</tr>
<tr>
<td>6: 4°C for infinity</td>
</tr>
</tbody>
</table>

#### 4.4.3.5. SAP Treatment

The SAP MM was prepared following the recipe mentioned in Table 4.4. Before loading 2µl into each well, the MM was placed on ice. The plate was then sealed with the same Polypropylene film.

The plates were centrifuge at 1000rpm for 1 minute at room temperature to remove any air bubbles, and after the volumes in each well was checked to ensure that they were the same,
the plate was run on SAP program in a Thermal Cycler (Veriti 384-Well Thermal Cycler, Applied Biosystems) (Table 4.5).

**Table 4.4 SAP Master Mix**

<table>
<thead>
<tr>
<th>Reagents</th>
<th>x1 reaction (in µl)</th>
<th>384 plate (with some room for pipetting errors) (in µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>1.53</td>
<td>612.00</td>
</tr>
<tr>
<td>SAP</td>
<td>0.30</td>
<td>120.00</td>
</tr>
<tr>
<td>SAP buffer</td>
<td>0.17</td>
<td>68.00</td>
</tr>
</tbody>
</table>

**Table 4.5 SAP Treatment Programme**

<table>
<thead>
<tr>
<th>SAP program</th>
<th>37°C for 30 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85°C for 10 min</td>
</tr>
<tr>
<td>2</td>
<td>95°C for 5 min</td>
</tr>
<tr>
<td>3</td>
<td>4°C for infinity</td>
</tr>
</tbody>
</table>

4.4.3.6. **Extension primer optimisation**

The concentration of each extension primer needed was adjusted in order to optimise the signal-to-noise ratio of MALDI-TOF MS. The mass of any analyte is inversely proportional to the signal-to-noise ratio, hence the analytes with a higher mass have a lower signal-to-noise ratio and this needs to be compensated for by improving their concentration. In the test plate, the extension primers were divided into 2 groups based on their mass (Table 4.6). The group with the smallest mass had the least amount of extension primers, and the group with the highest mass had the highest concentration of extension primers. Once the test plate had been completed, the concentration of extension primers was adjusted for the following plates based on the observed signal-to-noise ratio.
### Table 4.6 Division of the groups of SNPs based on their mass

<table>
<thead>
<tr>
<th>Assay 1</th>
<th>Assay 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3094509_W1</td>
<td>rs174537_W1</td>
</tr>
<tr>
<td>rs799923_W1</td>
<td>rs2946834_W1</td>
</tr>
<tr>
<td>rs1016990_W1</td>
<td>rs174575_W1</td>
</tr>
<tr>
<td>rs10486567_W1</td>
<td>rs4648310_W1</td>
</tr>
<tr>
<td>rs4794758_W1</td>
<td>rs7223952_W1</td>
</tr>
<tr>
<td>rs7405696_W1</td>
<td>rs12793759_W1</td>
</tr>
<tr>
<td>rs10896438_W1</td>
<td>rs2292884_W1</td>
</tr>
<tr>
<td>rs5275_W1</td>
<td>rs10993994_W1</td>
</tr>
<tr>
<td>rs1354774_W1</td>
<td>rs743572_W1</td>
</tr>
<tr>
<td>rs4962416_W1</td>
<td>rs17178655_W1</td>
</tr>
<tr>
<td>rs17576_W1</td>
<td>rs17552022_W1</td>
</tr>
<tr>
<td>rs6763931_W1</td>
<td>rs16987929_W1</td>
</tr>
<tr>
<td>rs2121875_W1</td>
<td>rs2242652_W1</td>
</tr>
<tr>
<td>rs5742657_W1</td>
<td>rs3918256_W1</td>
</tr>
<tr>
<td>rs174448_W1</td>
<td>rs1000778_W1</td>
</tr>
<tr>
<td>rs130067_W1</td>
<td>rs6162_W1</td>
</tr>
<tr>
<td>rs1535_W1</td>
<td></td>
</tr>
</tbody>
</table>

#### 4.4.3.7. iPLEX Extension

The iPLEX MM was prepared following the recipe mentioned in Table 4.7 and placed on ice. 2µl of iPLEX Extension MM was added to each well and then sealed with Polypropylene film. The plates were centrifuged at 1000rpm for 1 minute at room temperature to remove air bubbles. The volumes in each well were checked to be the same. The plate was then run on an extension program in a Thermal Cycler (Veriti 384-Well Thermal Cycler, Applied Biosystems) (Table 4.8).
Table 4.7 Extension Master Mix

<table>
<thead>
<tr>
<th>Reagent</th>
<th>x1 reaction (in µl)</th>
<th>384 plate (with some room for pipetting errors) (in µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>iPLEX buffer</td>
<td>0.2</td>
<td>90</td>
</tr>
<tr>
<td>iPLEX termination mix</td>
<td>0.2</td>
<td>90</td>
</tr>
<tr>
<td>Extension primer mix</td>
<td>1.559</td>
<td>701.55</td>
</tr>
<tr>
<td>iPLEX enzyme</td>
<td>0.04</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4.8 Extension PCR Programme

<table>
<thead>
<tr>
<th>Extension program</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84°C for 30 sec</td>
</tr>
<tr>
<td>2</td>
<td>94°C for 5 sec</td>
</tr>
<tr>
<td>3</td>
<td>52°C for 5 sec</td>
</tr>
<tr>
<td>4</td>
<td>80°C for 5 sec</td>
</tr>
<tr>
<td>5</td>
<td>Repeat step 3 - 4, for 5 times</td>
</tr>
<tr>
<td>6</td>
<td>Repeat step 2 - 4, for 40 times</td>
</tr>
<tr>
<td>7</td>
<td>72°C for 3 min</td>
</tr>
<tr>
<td>8</td>
<td>4°C for infinity</td>
</tr>
</tbody>
</table>

4.4.3.8. Resin clean-up

Clean Resin (Sequenom, Inc.) was air-dried and then spread over a 384-dimple plate with the resin in dimple-positions that corresponded the wells containing aliquots in the 384-well plate (with the processed samples). This was followed by smoothing over the 384-well plate using a scraper in order to remove the excess resin.

Care was taken to keep the 384-well plate with the sample faced down, on top of the dimple plate and then turned over to allow the resin to fall into the wells (with the samples).
After this, 25µl of Nano-pure water was added into each well. The 384-well plates were then centrifuged at 1500rpm for 2 minutes at room temperature to remove air bubbles.

The plates were rotated for 30 minutes on a rotator (Mini LabRoller™ Rotator, Labnet International, Inc.) to allow for proper mixing. Beyond it, the sample plates were centrifuged at 3000rpm for 15 minutes at room temperature and then stored at -20°C freezer before analysis by mass spectrometry at Sequenom Facility, Liggins Institute, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand.

4.4.3.9. Data analysis

Genotyped data was analyzed using the MassARRAY® Typer 4.0.2 Software (SEQUENOM 2006) and represented as call cluster plots (Figure 4.1). Each cluster represents one genotype. Uncertain genotypes that can’t be called by the Typer Software were determined manually using a Mass-Intensity spectrum plot. A homozygous allele had a single peak at the mass associated with the specific genotype allele. A heterozygous allele would have two peaks corresponding to two genotype alleles. The peaks should be approximately the same height as a heterozygous genotype should have both alleles in a 1:1 ratio.

In the “Traffic Lights” (Figure 4.1), the dark green dots represent successful run of the sample, the light green and yellow dots imply poor quality, and the red dots imply a failure (on purpose at times for negative controls such as water). Certain other examples of the “Traffic Lights” from our study are provided in the Appendices section.
Figure 4.1 A typical Call Cluster Plot. The call for a sample (well no. 1A) is called for here for the SNP rs6162.
4.5. Statistical Analysis

4.5.1. Lifestyle and demographics analyses

The age of the participants, and the tobacco smoking and the alcohol consumption status as risk for PCa (and, aggressive PCa) were analyzed using a Fisher exact test and Mann-Whitney U test was employed to identify the risk of high BMI and PCa (or, aggressive PCa).

4.5.2. SNP genotyping Data Analysis using PLINK

One of the most important advantages of using PLINK software for analysing SNP genotyping data is that it provides a comprehensive range of tools, including GWAS data manipulation, quality control, association studies (single SNP analysis and haplotype analysis), transmission disequilibrium testing (TDT), GWAS meta-analysis, epistasis, imputation and permutation. The results generated by PLINK can be used directly by other bioinformatic software (e.g. Haploview) for further in-depth analysis.

The PLINK input data file is stored in PED and MAP files. The genotype data stored in the PED file correlates to the corresponding MAP file. Therefore, any manual changes on these files were avoided, as it could render the data unusable by introducing an error.

Before going ahead with data analysis, the data obtained after the SEQUENOM® procedure was verified for being in compliance with HWE and for linkage disequilibrium using PLINK (Purcell, Neale et al. 2007; Vaidyanathan, Naidu et al. 2017).

4.5.2.1. Data analysis

A simple structured but effective method to detect population stratification, and implementation by PLINK (Vaidyanathan, Naidu et al. 2017) was used by us for analysing the SNP genotype data.

PLINK's clustering approach is based on the genome-wide average proportion of alleles shared identical-by-state between two individuals SNPs, i.e., pairing up the SNPs based on
genetic identity (Purcell, Neale et al. 2007). The IBS clustering is also used to test whether
the SNPs of two individuals belong to the same population. Following the stratification
analysis, we performed a standard case-control association test using a Cochran-Mantel-
Haenszel statistic (1 df (degree of freedom)) that tests for SNP-disease association conditional
on the clustering. This accounts for stratification effects, as has been reported in

To avoid the possibility of false positives with multiple SNP testing, statistical
significance can be restricted by the most conservative Bonferroni correction (BONF) along
with the less conservative Benjamini and Hochberg false discovery rate (BH-FDR) for
multiple testing corrections. However, as in most cases, the tested SNPs are already shown to
be statistically significantly associated with risk for PCa by other researchers, hence,
variations that demonstrated significant association to risk of PCa before BONF and BH-FDR
were also considered for further analyses by us (Vaidyanathan, Naidu et al. 2017).

4.5.2.2. Association analysis

To assess if the genes that were identified using a literature search, and the SNPs present
in them have an association with the samples begotten from a particular cohort, the data
generated by SNP genotyping is first assessed for association analysis (Purcell, Neale et al.
2007). To perform association analysis using PLINK we converted the genotype data in two
text files i.e. a pedigree file (.ped) and a map file (.map).

The command to perform an association analysis with multiple corrections is as follows:

\texttt{PLINK --file mydata --assoc --adjust}

The command generates an output file ‘PLINK.assoc’ which contains the following fields:

- \texttt{CHR} chromosome number,
- \texttt{SNP} SNP ID,
- \texttt{BP} base-pair position,
The above command outputs two files. One for the association (PLINK.assoc), which gives the results for the association analysis; and the other for the adjustment (PLINK.assoc.adjusted), which gives the results for the association analysis with adjustment for multiple factors.

The association analysis generates data to analyse and understand the effect of gene x environment interactions with regards risk of aggressive and non-aggressive PCa.

However, the data produced for the total list genes of interest that were considered before carrying out association analysis should further be checked for the “gene-only” effect as well in the cohort. This is carried out by an interaction analysis, after adjusting for the demographic factors which are statistically identified to have a significant association with the risk of PCa in our cohort.

4.5.2.3. Interaction analysis

It is so possible that certain diseases may be driven by external or, environmental factors rather than the genes having a main say. In order to identify the gene-specific cause of any particular human disease, it is important to analyse the data generated by SNP genotyping by reducing the statistical influence of the environmental factors on the disease (identified in 4.5.1).
To perform logistic interaction analysis using PLINK, we converted the genotype data in three binary files i.e. a BED file (.bed), a BIM file (.bim) and a FAM file (.fam) and a covariate file (.text) containing the covariates (i.e. the environmental factors).

The command to perform a logistic interaction analysis with multiple corrections is as follows:

```
PLINK --bfile mydata --logistic --genotypic --covar covariates.txt --covar-name C1,C2 --adjust
```

The aforementioned command outputs two files. One for the association logistics (PLINK.assoc.logistic), which gives the results for the association logistic analysis and the other for adjustment (PLINK.assoc.logistic.adjusted) which gives the results for the association logistic analysis with multiple adjustments.

This option generates the file PLINK.qassoc.gxe which contains the following fields:

- **CHR** Chromosome number
- **SNP** SNP identifier
- **NMISS1** Number of non-missing genotypes in first group (1)
- **BETA1** Regression coefficient in first group
- **SE1** Standard error of coefficient in first group
- **NMISS2** Number of non-missing genotypes in second group (2)
- **BETA2** Regression coefficient in second group
- **SE2** Standard error of coefficient in second group
- **Z_GXE** Z score, test for interaction
- **P_GXE** Asymptotic p-value for this test
4.5.2.4. Epistasis analysis

In order to define the SNP-SNP interaction on the progression of PCa, and identify any more possible SNP biomarkers, we performed an epistasis analysis. This was carried out by using PLINK employing the genotype data in two text files i.e. a pedigree file (.ped) and a map file (.map) (Purcell, Neale et al. 2007).

The command to perform an epistasis analysis is as follows:

\[ \text{PLINK --file mydata --epistasis} \]

This command outputs two files. One for the epistasis (PLINK.epi) which gives the results for the best epistasis analysis and the other for summary (PLINK.epi.cc.summary) which gives the results for all the epistasis analysis that were performed.

4.5.3. SNP genotyping Data Analysis using Artificial Neural Networks

In this stage, for input to the artificial neural network technique using MLP available in WEKA, an ARFF (Attribute-Relation File Format) file is created and then the first process of MLP is carried out. In this study, a feedforward artificial neural network accompanied by backpropagation learning model was applied together with input, output and hidden layers. Two different types of ANN-MLP analysis is conducted in this stage (Witten, Frank et al. 2016). One analysis is between total PCa (aggressive and non-aggressive) and control patient samples and the other analysis is between aggressive PCa and healthy control samples.

4.5.3.1. WEKA

WEKA, short for Waikato Environment for Knowledge Analysis, was formed by the University of Waikato in Hamilton, New Zealand (Hall, Frank et al. 2009). The WEKA platform is a compilation of data pre-processing tools and machine learning methods (Hall, Frank et al. 2009). It supplies substantial support for the complete course of exploratory data mining, involving developing the input data, assessing learning strategies statistically, and
visualizing the input data and the outcome of learning (Hall, Frank et al. 2009; Witten, Frank et al. 2016). WEKA contains a neural network classifier known as MLP that is trained by adopting the concept of back propagation. WEKA (Version 3.8.1) was used to carry out the neural network process by us.

4.5.3.2. ANN

ANNs are also known as connectionist models as the association weights signify the memory of the neural structure (Agatonovic-Kustrin and Beresford 2000). The ANN uses genetic algorithm (GA) to reduce the mean squared error of the training set data in the network. The fitness value is selected for maximum parameter to be choosing based on mean square error. The receiver operating characteristic curve (ROC) is used to measure the ratio of false positive and false negative values.

The training algorithms used are the LMAM (Levenberg-Marquardt with Adaptive Momentum) and OLMAM (Optimized Levenberg-Marquardt with Adaptive Momentum). The LM method is used for testing the non-linear squares. A feedforward neural network is used where the data for SNP genotyping is divided into four regions basal, parabasal, intermediate and superficial layer (Devi, Ravi et al. 2016).

4.5.3.3. MLP

In this study, the MLP neural network is employed to extract weights with higher accuracies. The option of the quantity of hidden layers is a crucial option to be contemplated when developing MLP-ANNs. It was demonstrated in 1989 by Cybenko that a neural network with a single hidden layer can estimate any continual functions (Cybenko 1989). Nonetheless, a neural network with multiple hidden layers targeting bigger problems was indicated by Popescu, et al., 2009 (Popescu, Balas et al. 2009). This can contribute to the training of the neural network determining for not many local minimums and depletion of the
network errors. For these reasons, a dual hidden layer perceptron network is chosen in this study.

The choice of an appropriate activation function for the neurons in the different layers of MLP is very crucial to the performance of the network. The linear activation function is generally used for the input neurons because it transmits the input dataset directly to the next layer with no transformation. The choice of activation function for the output layer neurons is a function of the problem being solved. In our case, which is a multiclass learning problem, we decided to select the sigmoid node function for the output layer neurons because it has the capability to handle either continuous values or $n$-class binary classification problems. The sigmoid node function is also chosen for the neurons in hidden layers because it is nonlinear and differentiable. Differentiability and nonlinearity are vital requirements for MLP training algorithms (Popescu, Balas et al. 2009).

MLP is a feedforward ANN model that maps sets of input data onto a set of suitable outputs and its most significant application has been in the emerging area of artificial intelligence. An MLP constitutes of several layers of nodes in a directed graph. MLP uses a supervised learning approach called backpropagation, as stated earlier, for training the network (Lo Brano, Ciulla et al. 2014).

The output of each node is called its "activation". Weight values are associated with each vector and node in the network, and these values constrain how input data are related to output data. Weight values associated with individual nodes are also known as biases. Weight values are determined by the iterative flow of training data through the network (i.e., weight values are established during a training phase in which the network learns how to identify particular classes by their typical input data characteristics).
4.5.3.4. **ARFF File Creation**

   For input to the ANN technique employing MLP, the data for the 623 patients (PCa versus Control) and 566 patients (Aggressive versus Control) was converted into two different ARFF files. In this step, for the first analysis, a first ARFF file was created which contained the SNP data for the 97 SNPs belonging to the 623 patient samples. The SNP data typically coded in biallelic codes such as A/T, C/C, G/C, etc. was converted into 16-bit binary codes, that is, using only the two numeric forms (0 and 1). For instance, the biallele A/A was converted into 1000000000000000 and A/T was converted into 0100000000000000 and so on. Overall, the ARFF file consisted of 623 instances, 1552 attributes, two datatypes (0 and 1) and two classes (PCa and healthy control). In total, there were 254 prostate cancer samples (197 aggressive samples and 57 non-aggressive samples) and 369 healthy control samples in this study for the first analysis. Also in this step, for the second analysis, a second ARFF file was created which contained the SNP data for the 97 SNPs belonging to the 566 patient samples. Similar conversion approach of converting biallelic codes into 16-bit binary codes as the first analysis was applied in this step. Overall, the ARFF file consisted of 566 instances, 1552 attributes, two datatypes (0 and 1) and two classes (PCa and Healthy Controls). In total, there were 197 aggressive PCa samples and 369 healthy control samples in this study for the second analysis.

4.5.3.5. **First Process of MLP**

   The first process of MLP (Stage-I) was applied on the two different datasets (i.e. on the 623 and 566 patient samples, respectively) for the 97 SNPs using the training set test option and cross-validation test option with 10 folds. The training set test option is where, 100% of samples are employed for training, whereas, the 10-fold cross-validation test option is where, 90% of samples are employed for training and 10% for cross-validation. The 10-fold cross-validation test option was employed in this study because its implementation has the potential
to resist over-fitting. The following parameters of MLP with two hidden layers; learning rate of 0.3; momentum of 0.2; training time of 150; seeding and validation set size as 0; validation threshold of 20 were chosen in this stage for both the analysis. The results were generated using the training set and 10-fold cross-validation test options of the first process of MLP.

4.5.3.6. Data Analysis– Stage II: Weights (SNPs) Extraction

In this stage, the weights were first extracted from the first process of MLP and their corresponding SNPs associated with that extracted weights were later extracted based on a set threshold value (more details below). Perceptrons can be trained by a basic learning algorithm that is commonly known as the delta rule. It computes the errors between the computed output data, and employs this to generate an alteration to the weights, therefore applying a framework of gradient descent. Nearly all perceptrons have outputs of 1 or -1 with a threshold of 0 and there is a little evidence that such network can be trained more rapidly than network generated from nodes with distinct activation and deactivation values.

In total, the first process of MLP for the first analysis generated 3104 weights. 1552 each for cancer and control samples, respectively. Maximum weights for the cancer and control samples were 5.62 and 4.59, respectively. Minimum weights for the cancer and control samples were -3.70 and -3.17, respectively. The average weights for the cancer and control samples were 0.0061 and 0.0019, respectively, whereas, the standard deviation values for the cancer and control samples were 0.383 and 0.309, respectively. A threshold value based on the standard deviation evaluations for the potential weights extraction was set to over one. Overall, based on the set threshold value of over one, for first analysis, 19 SNPs were extracted from the cancer weights and 14 SNPs were extracted from the control weights. All the duplicate SNPs were eliminated from this process. In total, 19 SNPs were removed and 14 SNPs were retained in this stage. Also, in total, the first process of MLP for the second analysis generated 3104 weights as the total number of SNPs remained the same in both the
analysis. Maximum weights for the cancer and control samples were 5.14 and 5.12, respectively. Minimum weights for the cancer and control samples were 2.67 and 2.86, respectively. The average weights for the cancer and control samples were 0.0027 and 0.0044, respectively, whereas, the standard deviation values for the cancer and control samples were 0.362 and 0.362, respectively. Overall, based on the set threshold value of over one, for second analysis, 20 SNPs were extracted from the cancer weights and 16 SNPs were extracted from the control weights. All the duplicate SNPs were eliminated from this process. In total, 22 SNPs were removed and 14 SNPs were retained in this stage.

4.5.3.7. Data Analysis– Stage III: Second MLP Process

In this stage, a second ARFF file is created using the extracted weights and then the second process of MLP is carried out on the extracted weights (SNPs). Two different types of ANN-MLP analysis on the similar datasets as Stage I is conducted in this stage. The main purpose of this stage is to see whether the overall cancer/control classification using the training set/10-fold cross validation strategy is still significantly good using the extracted SNPs from both analysis.

4.5.3.8. ARFF File Creation

The data for the 623 patients (PCa versus Control) and 566 patients (Aggressive versus Control) was converted into two different ARFF files for input to the second process of MLP. In this step, for the first analysis, a first ARFF file was created which contained the SNP data for the 14 extracted SNPs belonging to the 623 patient samples. As previously stated in section 3.4, the SNP data was converted into 16-bit binary codes. Overall, the ARFF file consisted of 623 instances, 224 attributes, two datatypes (0 and 1) and two classes (PCa and Healthy Controls). Also in this step, for the second analysis, a second ARFF file was created which contained the SNP data for the 14 SNPs belonging to the 566 patient samples. Overall,
the ARFF file consisted of 566 instances, 224 attributes, two datatypes (0 and 1) and two classes (PCa and Healthy Controls).

4.5.3.9. Second Process of MLP

The second process of MLP (Stage-III) was applied on the two different datasets (i.e. on the 623 and 566 patient samples, respectively) for the 14 extracted SNPs using the training set test option and cross-validation test option with 10 folds. The following parameters of MLP with two hidden layers; learning rate of 0.3; momentum of 0.2; training time of 150; seeding and validation set size as 0; validation threshold of 20 were chosen in this stage for both the analysis. The results of the second process of MLP (Stage-III) was also generated using the training set and 10-fold cross-validation test options for the first and second analysis.
### 4.6. Cell line studies

The expression of the proteins produced by genes of interest, identified after comparing the results obtained by using PLINK and ANN analyses, was verified in PCa cell lines: LNCaP, DU145 and PC3 as well as a A431- skin cancer cell line and two ovarian cancer cell lines- SKOV3, and OVCAR3.

The cells were incubated in a SANYO (MCO-18M) cell culture incubator at 37°C in an atmosphere of 5% Carbon Dioxide (CO₂). Cell media contained all the nutrients and growth factors required to sustain cell growth. The cells were harvested when they reached close to 100% confluency; otherwise, cell media was renewed weekly to maintain cell growth.

### 4.6.1. Reagents and Equipment

The reagents used for cell culturing are mentioned in Table 4.9 and Table 4.10, and the equipment used are mentioned in Table 4.11.

<table>
<thead>
<tr>
<th>Media</th>
<th>Components</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCS P+S+G+ (stored at 4°C)</td>
<td>Fetal calf serum (FCS) (stored at 4°C) 10%</td>
<td>Morgate Biotech, Hamilton, New Zealand</td>
</tr>
<tr>
<td></td>
<td>Penicillin G (100U/ml)</td>
<td>Sigma Chemical Company, St Louis, MO, USA.</td>
</tr>
<tr>
<td></td>
<td>Streptomyacin sulphate (100U/ml)</td>
<td>Sigma Chemical Company, St Louis, MO, USA.</td>
</tr>
<tr>
<td></td>
<td>0.3g/litre Glutamine (G)</td>
<td>Ambion Ltd. Auckland, New Zealand</td>
</tr>
<tr>
<td></td>
<td>RPMI-1640 Medium</td>
<td>Gibco</td>
</tr>
<tr>
<td></td>
<td>DMEM, high glucose</td>
<td>Gibco</td>
</tr>
<tr>
<td></td>
<td>Antibiotic-Antimycotic</td>
<td>Thermo Fisher</td>
</tr>
</tbody>
</table>
Table 4.10 Phosphate Buffered Saline (PBS)

<table>
<thead>
<tr>
<th>PBS</th>
<th>Components per litre? and Suppliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjust to PH=7.4</td>
<td>0.20g of Potassium Chloride (KCl)</td>
</tr>
<tr>
<td></td>
<td>0.24g of Potassium Dihydrogen Phosphate (KH₂PO₄)</td>
</tr>
<tr>
<td></td>
<td>1.44g of Disodium hydrogen orthophosphate (Na₂HPO₄)</td>
</tr>
<tr>
<td></td>
<td>8.00g of Sodium chloride (NaCl)</td>
</tr>
<tr>
<td></td>
<td>1.00L of Mili-Q Water</td>
</tr>
</tbody>
</table>

Table 4.11 Equipment

<table>
<thead>
<tr>
<th>Product</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>HERA Safe KSP Biological Safety hood</td>
<td>Thermo Fisher Scientific</td>
</tr>
<tr>
<td>Flasks</td>
<td>Corning Inc.</td>
</tr>
<tr>
<td>Microscope</td>
<td>Millennium Science (Evos core xl)</td>
</tr>
<tr>
<td>CO₂ Incubator</td>
<td>SANYO (MCO-18M)</td>
</tr>
</tbody>
</table>

4.6.2. Protocol

4.6.2.1. Thawing cell lines

Cells preserved in cryovials containing frozen cell suspension were rapidly thawed in a 37°C water bath until most cells were thawed (1 to 2 minutes). Outside of the vial was disinfected with 70% ethanol and the thawed volume of cells was transferred, using aseptic technique, into a T-25 tissue culture flask containing 20 mL of growth medium. The cells
were cultured at 37°C in a humidified environment with 5% CO₂. After 24 hours, once the cells had adhered to the flask, the medium was replaced with fresh culture medium to remove DMSO-containing freezing medium.

4.6.2.2. Subculture procedure

Cell growth medium was removed from the flask and discarded once cells were at confluency. The cells were rinsed twice with warm PBS (5 mL) in order to get rid of growth medium followed by 1 mL of Trypsin-EDTA solution to detach cells from flask surface. The flask was tilted to ensure all cells were exposed to the Trypsin-EDTA and then incubated at 37°C for 3 to 4 minutes. The cells were observed under the microscope (using the 20x magnification) to ensure that the cells have detached. 3 mL of growth medium was then added to stop trypsinisation and the cells were mixed gently and centrifuged at 1000 rpm for 5 min. The cell pellet was mixed with 5 mL of growth medium and the cell suspension was added to a new flask containing fresh culture medium.

4.6.2.3. Cryopreservation of cells

Cells were cryopreserved to maintain their viability and to maintain cells stocks at low subculture number. Prior to cryopreservation cells were checked for optimal growth without any contamination. Freezing medium was made up with DMSO, growth medium and fetal calf serum at the ratio 1:3:6. Cells were trypsinized as in the subculture procedure and were then suspended in growth medium. After centrifuging for 5 minutes at 1000 rpm at room temperature, cell pellets were resuspended in an appropriate volume of freezing medium and aliquoted at 1x10⁶ cells per 1 mL per cryovial. After being clearly labelled, these cryovials were transferred into Mr. Frosty™ Freezing Container and were put into -80°C freezer. After approximately 24 hours, cryovials were transferred into liquid nitrogen for long term storage.
4.7. Western Blot

4.7.1. Introduction

Western blot, also known as protein immunoblotting, is a widely-adopted methodology for detecting the amount of a specific protein in a certain sample. Firstly, polyacrylamide gel electrophoresis is used to separate proteins in the lysate of cells or tissues. Then proteins are immobilized onto a sheet of Polyvinylidene difluoride (PVDF) membrane. The procedure of transferring protein to a sheet of membrane was first described by Harry Towbin in 1979, and the idea came from “Southern blot” which transfers and detects DNA samples. Ultimately, the immobilized proteins can be detected by labelled antibodies using chemiluminescent western blot detection, which is a highly sensitive alternative to isotopic detection of proteins bound to blotting membranes.

4.7.2. Reagents and Equipment

<table>
<thead>
<tr>
<th>Product name</th>
<th>Components/Suppliers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell lysis buffer</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 mM Tris (Invitrogen) pH7.5</td>
</tr>
<tr>
<td></td>
<td>150 mM sodium chloride</td>
</tr>
<tr>
<td></td>
<td>1 mM EDTA</td>
</tr>
<tr>
<td></td>
<td>1 mM EGTA</td>
</tr>
<tr>
<td></td>
<td>1% Triton X1000</td>
</tr>
<tr>
<td></td>
<td>2.5 mM sodium pyrophosphate</td>
</tr>
<tr>
<td></td>
<td>1 mM β-glycerol phosphate</td>
</tr>
<tr>
<td></td>
<td>1 mM vanadate</td>
</tr>
<tr>
<td></td>
<td>100 mM NaF</td>
</tr>
<tr>
<td><strong>Protease inhibitors</strong></td>
<td></td>
</tr>
<tr>
<td>Protease inhibitors (added freshly before use)</td>
<td>9.4 μM Leupeptin</td>
</tr>
<tr>
<td></td>
<td>1.5 μM Pepstatin A</td>
</tr>
<tr>
<td></td>
<td>1 mM AEBSF</td>
</tr>
<tr>
<td></td>
<td>0.6 μM Aprotinin</td>
</tr>
<tr>
<td></td>
<td>30 μM ALLN</td>
</tr>
<tr>
<td></td>
<td>1 mM DTT</td>
</tr>
<tr>
<td><strong>Loading buffer</strong></td>
<td></td>
</tr>
<tr>
<td>5x SDS Loading blue</td>
<td>4% SDS</td>
</tr>
<tr>
<td></td>
<td>0.2 M Tris (pH 6.8)</td>
</tr>
<tr>
<td></td>
<td>50% Glycerol (100%)</td>
</tr>
<tr>
<td></td>
<td>200 mM DTT</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>1% bromophenol blue</td>
</tr>
</tbody>
</table>

**Running buffer**

<table>
<thead>
<tr>
<th>SDS-PAGE GEL running buffer</th>
<th>25 mM Tris</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250 mM Glycine</td>
</tr>
<tr>
<td></td>
<td>0.1% SDS</td>
</tr>
</tbody>
</table>

**Transfer buffer**

<table>
<thead>
<tr>
<th>Membrane side buffer</th>
<th>300 mM Tris</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>300 mM Glycine</td>
</tr>
<tr>
<td></td>
<td>20% ethanol</td>
</tr>
<tr>
<td>Gel side buffer</td>
<td>300 mM Tris</td>
</tr>
<tr>
<td></td>
<td>300 mM Glycine</td>
</tr>
<tr>
<td></td>
<td>0.1% SDS</td>
</tr>
</tbody>
</table>

**Blocking buffer**

| BSA blocking buffer          | 3 g BSA powder in 100 ml of TBST                       |

**Antibodies**

<table>
<thead>
<tr>
<th>MYEOV</th>
<th>Cell signaling technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALK</td>
<td>Symansis/ In house</td>
</tr>
<tr>
<td>PPAR-γ</td>
<td>Symansis/ In house</td>
</tr>
<tr>
<td>β-catenin (rabbit)</td>
<td>Symansis/ In house</td>
</tr>
</tbody>
</table>

**ECL substrates**

<table>
<thead>
<tr>
<th>Clarity™ Western ECL Blotting Substrate</th>
<th>Bio-rad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Lightning Ultra</td>
<td>PerkinElmer</td>
</tr>
</tbody>
</table>

**Stripping buffer**

<table>
<thead>
<tr>
<th>Stripping buffer</th>
<th>100 mM β-mercaptoethanol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2% SDS (w/v)</td>
</tr>
<tr>
<td></td>
<td>62.5 mM Tris-HCl, pH 6.7</td>
</tr>
</tbody>
</table>

**Membrane washing buffer**

<table>
<thead>
<tr>
<th>TBST</th>
<th>10 mM Tris</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 mM NaCl</td>
</tr>
<tr>
<td></td>
<td>1.34 mM KCl</td>
</tr>
<tr>
<td></td>
<td>0.1% Tween-20 (v/v)</td>
</tr>
<tr>
<td>TBS</td>
<td>10 mM Tris</td>
</tr>
<tr>
<td></td>
<td>10 mM NaCl</td>
</tr>
<tr>
<td></td>
<td>1.34M KCl</td>
</tr>
</tbody>
</table>

**Equipment**

<table>
<thead>
<tr>
<th>Multi-Well plates</th>
<th>Corning. Inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microfuge 22R Centrifuge</td>
<td>Beckman Coulter</td>
</tr>
<tr>
<td>Gel electrophoresis equipment:</td>
<td>PowerPac™ Basic Power Supply Mini-PROTEAN® Tetra Cell Systems (BIO-RAD)</td>
</tr>
<tr>
<td>Trans-Blot® Turbo™ Transfer System</td>
<td>BIO-RAD</td>
</tr>
<tr>
<td>VSR-50 Laboratory Platform Rocker</td>
<td>Lab plus series, PRO scientific</td>
</tr>
<tr>
<td>Water bath:</td>
<td>Grant</td>
</tr>
<tr>
<td>Luminescent image analyser</td>
<td>Fujifilm LAS 4000</td>
</tr>
<tr>
<td>Synergy 2 Multi-Mode Reader</td>
<td>BioTek</td>
</tr>
</tbody>
</table>
4.7.3. Protocol

4.7.3.1. Protein lysate preparation (taking a 6-well plate as an example)

At the end of designed treatments, cells were rinsed twice with cold PBS and lysed by the addition of 200 uL of lysis buffer (containing protease inhibitors) per well. After being incubated on ice for 10 minutes, wells were scraped and then the lysates were transferred into pre-chilled tubes. After a 5-minute incubation on ice, lysate tubes were briefly vortexed and were centrifuged at 4°C for 10 minutes at 13000xg. The supernatant from the wells was removed and used as the lysate samples for western blotting.

4.7.3.2. Protein quantification

Protein concentrations of lysate samples were measured in a 96 well plate by Pierce™ BCA Protein Assay Kit (ThermoFisher cat# 23225). 2 mg/mL of BCA stock solution was used for standard curve. 2 uL of each sample and standard curve point were measured in duplicate and 2 uL of lysis buffer alone was also measured to determine background signal. 200 μL of BCA reagent mixture (reagent A to reagent B ratio=50:1) was added into each well and the plate was covered and incubated at 60°C for 15-30 minutes. After the plate cooled down to room temperature, the absorbance at 562nm was measured using a plate reader (BioTek Synergy 2 Multi-Mode Reader). The concentration of each unknown sample was then calculated using the standard curve.

4.7.3.3. Running PAGE gel and transferring

The appropriate volume of protein required was mixed with 6x loading buffer and loaded into the wells of the SDS-PAGE gel, along with the appropriate molecular weight markers. The gel was run at 80-100V for approximately 30 minutes to allow proteins to enter the stacking gel and then run at 100-120V for around 60 minutes until the dye front was near the bottom of the resolving gel. Membrane side transfer buffer and gel side transfer buffer were
prepared and PVDF membranes were first activated in methanol and then equilibrated with membrane side buffer, together with membrane side filter paper. Another set of filter paper was equilibrated with gel side buffer. The transfer sandwich was set up in the Trans-Blot® Turbo™ Transfer cassette and the semi-dry transfer system was then run at 25V, 2.5A for 12 minutes.

4.7.3.4. Antibody blotting

After the transfer, the membrane was blocked in 3% BSA for 1 hour at room temperature. Antibodies of interest were diluted in 3% BSA according to manufacturer’s instructions. Subsequently, membranes were incubated in diluted primary antibodies at 4°C overnight with gentle shaking. At the end of the incubation, primary antibodies were recycled and the membranes were washed three times in TBST for 5 minutes each time with gentle shaking. Membranes were then incubated in corresponding secondary antibodies diluted 1:10,000 in 3% BSA at room temperature for 1 hour. At the end of the incubation, antibody solution was removed and the membranes were washed three times in TBST for 5 minutes each time with gentle shaking. After the last wash with TBST, membranes were incubated with ECL Blotting Substrates (Bio-Rad) at room temperature for 5 minutes and then the images were captured using the LAS 4000 Imager.

4.7.3.5. Stripping and re-probe

After imaging, the membranes can still be used for probing with further antibodies, such as for the loading controls or total protein measurement after detection of phosphoproteins. To strip membranes of antibody, they were soaked in stripping buffer and incubated at 55°C for 20 minutes. After times three of 15 minutes washing with TBS, membranes were ready for a new round of blocking and antibody incubation, if needed.
“If I can't make it through one door, I'll go through another door- or I'll make a door. Something terrific will come no matter how dark the present.”

- Rabindranath Tagore
Chapter 5. SNP association with risk of prostate cancer

Chapter preamble

SNP genotyping is a measure of the genetic variations of the polymorphisms across various members of the same species in order to identify the risk association of the alleles and the impact of environmental factors on the same with regards the prognosis of any disease of interest. It is a form of genotyping, which is the measurement of more general genetic variation. SNPs, being one of the most common types of genetic variations and making-up more than around 90% of the entire human genetic variations have been extensively studied for functional and structural relationships between genotype and phenotype (disease- PCa in this case).

SNP usually comprise of two alleles, where the rare allele frequency is >1%. SNPs are found to be involved in the etiology of many human diseases and are thus becoming of much interest in molecular and pharmacological studies.

It is well established that GWAS do not verify for exonic, intronic and/or intergenic SNPs and the impact of external and/or environmental factors on the SNPs identified across considerably large populations. The effect of exonic SNPs is much clearer as they can affect the protein expression level(s) and/or alteration in the amino acids coding for certain protein, but intronic SNPs are much more challenging. They may affect splice variants and form isoforms, thereby lead to variation in chromosomal structures, and yet may not be detected in various protein and/or gene expression analyses. Various case-control studies have been carried out to identify the effect of gene x environment interaction(s).

Section 5.1 has been presented in the form of a published manuscript considering the gene x environment interactions and the risk of aggressive PCa with regards New Zealand perspective.
Environmental factors and risk of aggressive prostate cancer among a population of New Zealand men- a genome-wide association analysis
5.1. Environmental factors and risk of aggressive prostate cancer among a population of New Zealand men- a genome-wide association analysis

Abstract: Prostate cancer is one of the most significant male health concerns worldwide. Numerous researchers carrying out molecular diagnostics have indicated that genetic interactions with biological and behavioral factors play an important role in the overall risk and prognosis of this disease. Single nucleotide polymorphisms (SNPs) are increasingly becoming strong biomarker candidates to identify susceptibility to prostate cancer. We carried out a gene x environment interaction analysis linked to aggressive and non-aggressive prostate cancer (PCa) with a number of SNPs. By this method, we identified the susceptible alleles in a New Zealand population, and examined the interaction with environmental factors. We have identified a number of SNPs that have risk associations both with and without environmental interaction. These indicate that certain SNPs have been associated with disease vulnerability based on behavioral factors. The list of the genes with SNPs identified as risk of PCa in a New Zealand population is mentioned in Figure 5.1.

Figure 5.1 Pictorial Abstract
5.1.1. Introduction

Prostate cancer (PCa) is one of the most significant male health concerns worldwide, due to its high prevalence and a risk of around 1 in 6 patients developing the aggressive form of this disease (Cooperberg, Vickers et al. 2010). Understanding of the risk of the disease becoming aggressive is important for the appropriate management of PCa (Scher and Heller 2000). PCa may start as an indolent disease of the prostate gland, followed by non-castrate increase in prostate-specific antigen (PSA), followed by a non-castrate metastatic stage and finally progression to the lethal castration-resistant, aggressive disease (Chang, Pao et al. 2013).

In 2002, PCa was identified as the third-most prevalent cancer among adult males worldwide (Quinn and Babb 2002). Ten years after, PCa, is the second-most common cancer among men and the sixth leading cause of death due to cancer among males worldwide (Center, Jemal et al. 2012). The highest recorded rate of men with PCa, relative to the population of healthy men, is observed in the Oceania region (Jemal, Bray et al. 2011; Karunasinghe, Han et al. 2012).

Although a heredity aspect is known for PCa (Bratt 2002), studies have also shown that genetic interactions between biological and behavioral factors play an important role in the overall risk and prognosis of PCa (Schaid 2004; Karunasinghe, Han et al. 2012; Karunasinghe, Lange et al. 2013). Since decisions of clinical management are also based on identifying the risk of aggressive PCa (Chang, Pao et al. 2013), extensive research is being carried out worldwide to identify the role played by single nucleotide polymorphisms (SNPs), their association in the development and progression of PCa and their potential use as biomarkers (Orozco, Goh et al. 2013; Van den Broeck, Joniau et al. 2014). SNPs are increasingly becoming strong biomarker candidates to identify PCa susceptibility (Tao, Wang et al. 2012; Orozco, Goh et al. 2013; Van den Broeck, Joniau et al. 2014). Due to the
increasing number of cases of PCa, identifying one or more biomarkers for early detection and proper management of PCa to avoid disease progression into the aggressive state is becoming an urgent priority internationally (Velonas, Woo et al. 2013).

It is recognized that one form of defense against cancer development involves a series of genes whose role is to metabolize and excrete potentially toxic compounds and to repair subtle mistakes in DNA such as the mismatch repair genes. Particular environmental exposures can exacerbate the genetic influence on PCa through gene x environment interaction (Brennan 2002).

This paper explores the use of SNP genotypes as biomarkers for aggressive PCa. Here we present the data obtained following the genotype analysis of 136 SNPs, located in 66 genes and 10 chromosomal locations using SEQUENOM MassArray technology and the TaqMan SNP genotyping procedure. The cohort includes New Zealand men (of European descent) with different grades/stages of PCa, and age matched male controls. We seek to detect the association of SNPs with both aggressive and non-aggressive disease as well as the influence of external factors in risk modification. This, we believe, is the first such study on genetic and environmental risk association and interaction analysis leading to aggressive PCa in a New Zealand cohort.

5.1.2. Materials and Methods

5.1.2.1. Study population

Patients with proven diagnosis for PCa from the Auckland Regional Urology Registry (Auckland, Middlemore, and North Shore hospitals), and private practices in the Waikato region, New Zealand were invited to take part in this study between 2006-2014 (ethics reference NTY05/06/037). Patient recruitment was carried out with their informed consent. Initially, patients were recruited within one year of diagnosis, if they had not undergone any
treatment for PCa. In 2008, the criterion was relaxed to include all patients with malignancies, regardless of treatment but within one year of diagnosis. In September 2010, the time frame for recruitment was removed altogether. From those that took part in this study, a total of 197 men were identified with aggressive PCa, and 57 with non-aggressive PCa. Additionally, 369 healthy males with no symptoms of PCa took part in our studies (ethics reference NTY/06/07/AM04), with no signs of the established PCa symptoms such as the urge to urinate frequently with increased urination during night time, sudden urge to urinate, and difficulty in maintaining a steady stream of urine, hematuria and dysuria were considered as “healthy controls” (Hodgson, Obertova et al. 2012). The men with a self-reported European ancestry were included in the study. The age at which PCa was diagnosed in patients and the age at recruitment for the healthy controls were considered as age for analysis. It varied from 40-81 years across the groups of men recruited for this study.

5.1.2.2. Data collection for demography, and lifestyle details

Each patient completed a demographic and lifestyle questionnaire at entry into the study. The questionnaire included details about the individual’s history of smoking tobacco, consumption of alcohol, body mass index (BMI) at time of recruitment, and age at diagnosis of PCa. Current smokers and individuals who reported a history of smoking tobacco were jointly considered as smokers.

The clinical history of each patient was extracted from hospital databases and transferred to a central study database.

5.1.2.3. Collection and processing of blood samples

Blood samples from each volunteer were collected in Vacutainer® tubes (Becton Dickinson) containing Ethylene Diamine Tetra Acetic acid (EDTA). An aliquot of the sample collected was then used for genomic DNA extraction. Each patient’s DNA was extracted
using a QIAamp genomic DNA kit (Qiagen, Hilden, Germany) following the manufacturers’
protocol with the aid of a fully automated QIAcube (Qiagen, Hilden, Germany). The DNA
samples were diluted to 5.0ng/μl as per requirement of the SEQUENOM MassARRAY
iPLEX® assay protocol.

5.1.2.4. Selection of SNPs

A total of 136 SNPs, located in 66 genes and some undefined chromosomal locations
(Supplementary Table 1) were identified by a literature search of the published genome-wide
association studies (GWAS) for both PCa and its aggressive form.

The PubMed database was screened for research articles only that provided full text in
English using key terms “prostate cancer” and/or “aggressive prostate cancer”, along with
“SNP genotyping sequenom massarray iplex”, and/or “genome-wide association studies”.
Only research articles published on or after the year 2000 till the year 2013 were considered
for this study to maintain the current trend of research at that time. Since the sample
population for this research is of European ancestry, attention was given to research papers
which reported statistically significant findings among patients with such backgrounds.

The research team has also used their own knowledge and discretion regarding which
SNPs to be genotyped using SEQUENOM MassARRAY iPLEX® assay, or the TaqMan®
SNP genotyping assay. Various metabolic pathways such as selenium metabolism and
androgen metabolism pathway were included, for instance, as was, screening for
chromosomal regions such as 8q24, 10q11, 17q12, and Xp11 all known for PCa risk
association. Additionally certain putative oncogenes such as MYEOV (Myeloma
Overexpressed), and DNA mismatch repair genes such as MLH1 (MutL homolog 1); as well
as certain genes with an established role in the androgen metabolism pathway (Karunasinghe,
Han et al. 2012; Karunasinghe, Han et al. 2012; Bishop, Han et al. 2016) were also
considered in this assessment.
Our approach uses genome wide association analysis (GWAA) to provide a comprehensive evaluation of multiple genes with polymorphisms that interact in the same pathway/s, in line with proposals made by Hsing et al., (2008) (Hsing, Chu et al. 2008) and Kwon et al., (2012) (Kwon, Holt et al. 2012), who claim that studies on GWAA are a better way to study multifactorial diseases such as PCa. Also, Hsing et al., (2008), have suggested that research pertaining to the AR pathway should include co-regulators along with important genes involved in androgen metabolism and biosynthesis (Hsing, Chu et al. 2008). A similar approach was followed in our selection of genes with the SNPs studied below.

5.1.2.5. Genotyping of candidate genes

5.1.2.5.a. SNP genotyping by Sequenom multiplexing

Genotyping for the candidate SNPs was carried out in the Auckland UniServices Sequenom Facility at Liggins Institute, Auckland, and AgResearch Limited, Mosgiel, New Zealand, using custom-designed multiplex gene panel and iPLEX chemistry. Genotype calling was performed using the standard post-processing calling parameters in SEQUENOM Type 4.0 software. Each 384-well plate prepared for genotyping contained known HAPMAP control samples, negative controls (Water) and repeats of samples used in different locations in the 384-well plate for validation of the genotyping procedure.

5.1.2.5.b. SNP genotyping by TaqMan® assay

SNP genotyping using TaqMan® SNP genotyping (Applied Biosystems (ABI)) was carried out on a panel of genes. The primers used were either obtained pre-designed from ABI or were custom-made using Assay-by-Design service by ABI (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2012; Bishop, Han et al. 2016). The protocol provided by the manufacturers (ABI) was used to carry out the SNP genotyping (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2012; Bishop, Han et al. 2016).
5.1.3. Collection of clinical characteristics of patients

The total serum PSA level and Gleason score of the tumour at biopsy or post-surgery or both were obtained from the patients’ clinical records at the time of recruitment or subsequent to recruitment. Post-surgery Gleason score was given priority over biopsy Gleason score. Staging data was collected from clinical records during the years 2013-2014, but was not available for all the patients.

5.1.3.1. Definition of aggressiveness

The aggressiveness of PCa for this study is based on the classification used by the American Joint Committee on Cancer, Prostate Cancer Staging, 7th Edition (Wong, Yip et al. 2014) and first proposed by D’Amico et al. (1998) for defining high-risk or aggressive PCa as clinical T stage ≥T2c, or Gleason score ≥8, or serum PSA level ≥20ng/ml (D’Amico, Whittington et al. 1998).

5.1.3.2. Statistical analysis

5.1.3.2.a. SNP data cleaning

A total of 39 SNPs were removed from the list before data analysis after checking the genotype for compliance with Hardy Weinberg Equilibrium (HWE) and in order to avoid linkage disequilibrium using PLINK (Purcell, Neale et al. 2007). 13 SNPs were removed for not complying with HWE and an additional 26 SNPs were removed for being in linkage. The entire list of the SNPs removed from the final analyses is colour-coded in the Supplementary table 1. We therefore analyzed the data using a reduced list of 97 SNPs.

Analysis of the data for SNP association with PCa based on aggressiveness and gene x environment interaction for risk of PCa were both carried out using PLINK - a tool set for whole genome association(Purcell, Neale et al. 2007). Correction for multiple testing was
applied to significance value where appropriate otherwise, statistical significance was set at 
\( p \leq 0.05 \) (Balding 2006).

### 5.1.3.2.b. Gene-environment interaction for risk of PCa

A structured association approach similar to the one proposed by Arya et al., in 2009 (Arya, Hare et al. 2009), working on rheumatoid arthritis, was followed. This provides a simple but powerful method, to detect population stratification, and is implemented in the PLINK (Pritchard, Stephens et al. 2000; Purcell, Neale et al. 2007; Arya, Hare et al. 2009).

PLINK’s clustering approach is based on the genome-wide average proportion of alleles shared identical-by-state (IBS) between two individuals SNPs, i.e., pairing up the SNPs based on genetic identity (Purcell, Neale et al. 2007). The IBS clustering is used to test whether the SNPs of two individuals belong to the same population. Following the stratification analysis, we performed a standard case-control association test using a Cochran-Mantel-Haenszel statistic (1 df (degree of freedom)) that tests for SNP-disease association conditional on the clustering. This accounts for stratification effects, as has been reported by Arya et al. (2009) (Arya, Hare et al. 2009).

To avoid the possibility of false positives with multiple SNP testing, statistical significance was restricted by the most conservative Bonferroni correction (BONF) along with the less conservative Benjamini and Hochberg false discovery rate (BH-FDR) for multiple testing corrections. The complete set of results is provided in Supplementary Table 2. However, as the tested SNPs were already shown statistically to be significantly associated with PCa risk by other researchers (Kote-Jarai, Olama et al.; Eeles, Olama et al. 2013; Karunasinghe, Lange et al. 2013), variations that demonstrated significant association to risk of PCa before BONF and BH-FDR were considered for discussion in our study.
5.1.3.2.c. **Significance of demographic factors**

The variation of lifestyle characteristics between pathology was tested using the Fisher exact test.

The BMI data were not normally distributed; therefore, comparisons between pathologies were carried-out using the Mann Whitney test.

5.1.3.2.d. **SNP association analysis for risk of Prostate Cancer**

The covariates included the three environmental factors of- BMI record, tobacco smoking and alcohol consumption history. Also, three genotypic models: additive (ADD), dominant deviation (DOMDEV), and general _2 df joint test of both additive and dominant deviation (GENO_2DF) were generated. An additive model represents the additive effects of SNPs i.e., the effect of each additional minor allele as represented by the direction of the regression coefficient. For example, a positive regression coefficient indicates that the minor allele increases risk. A DOMDEV model represents a separate test of the dominance component, and a general model represents the joint test of both ADD and DOMDEV components. However, in contrast to a dominance model, ADD refers to a variable coded in such a way (0, 1, 0 for three genotypes AA, Aa, aa) that it represents the dominance deviation from additivity without specifying whether a particular allele is dominant or recessive. Effects of genotype × environmental (BMI, tobacco smoking and alcohol consumption) interactions was tested and eventually corrected for.

5.1.4. Results

5.1.4.1. **Gene x environment effects and prostate cancer risk in a New Zealand population**

The tables show the results of the statistically significant SNPs associated between non-aggressive PCa and healthy controls (Table 5.1), between patients with aggressive and non-
aggressive PCa (Table 5.2), and patients with aggressive PCa and healthy controls (Table 5.3), all assessed before the use of BONF.

The results obtained after a detailed analysis of the association of SNPs as risk of PCa can be broadly classified into two categories- expected, and unique with respect to the results reported by other groups.

Certain SNPs present near genes such as *MYEOV* - a putative oncogene (Table 5.3); *TLR4* (*Toll-like receptor 4*) (Tables 5.1 and 5.2) and *MMP9* (*Matrix metallopeptidase 9*) (Table 5.2)- involved with the inflammatory pathway; and *KLK3* (*Kallikrein-3*) (Table 5.2 and 5.3) and *MSMB* (*Microseminoprotein Beta*) (Table 5.3)- both involved in the androgen metabolism pathway were understandably identified as statistically significant in our study, due to their proven risk association to PCa.

The number of SNPs present near various genes associated with obesity and diabetes mellitus such as *FADS2* (*Fatty acid desaturase 2*) (Table 5.3), *LEP* (*Leptin*) (Tables 5.2 and 5.3), *PPAR-γ* (*Peroxisome Proliferator-Activated Receptor gamma*) (Table 5.3) were associated with the risk of aggressive PCa vs healthy controls as recorded in our analysis.

**Table 5.1 Statistically significant case-control SNP association between patients with non-aggressive prostate cancer and healthy controls**

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene location</th>
<th>Tested allele</th>
<th>Odds Ratio (95% CI)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2292884</td>
<td>MLPH</td>
<td>2q37.2</td>
<td>G</td>
<td>1.774 (1.074-2.931)</td>
<td>0.02375</td>
</tr>
<tr>
<td>2</td>
<td>rs4965373</td>
<td>SEPS1</td>
<td>15q26.3</td>
<td>A</td>
<td>1.801 (1.074-3.022)</td>
<td>0.02413</td>
</tr>
<tr>
<td>3</td>
<td>rs11536889</td>
<td>TLR4</td>
<td>9q33.1</td>
<td>C</td>
<td>2.198 (1.075-4.494)</td>
<td>0.02727</td>
</tr>
<tr>
<td>4</td>
<td>rs3735035</td>
<td>PODXL</td>
<td>7q32</td>
<td>T</td>
<td>0.636 (4168-0.9707)</td>
<td>0.03493</td>
</tr>
</tbody>
</table>
Table 5.2 Statistically significant case-case SNP association between patients with aggressive prostate cancer and non-aggressive prostate cancer

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene location</th>
<th>Tested allele</th>
<th>Odds Ratio (95% CI)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs632148</td>
<td>SRD5A2</td>
<td>2p23.1</td>
<td>C</td>
<td>1.799 (1.105-2.929)</td>
<td>0.01731</td>
</tr>
<tr>
<td>2</td>
<td>rs887391</td>
<td>SLC26A6</td>
<td>19q13</td>
<td>C</td>
<td>1.793 (1.089-2.952)</td>
<td>0.02063</td>
</tr>
<tr>
<td>3</td>
<td>rs11536889</td>
<td>TLR4</td>
<td>9q33.1</td>
<td>C</td>
<td>0.4341 (0.2084-0.9042)</td>
<td>0.02251</td>
</tr>
<tr>
<td>4</td>
<td>rs2292884</td>
<td>MLPH</td>
<td>2q37.2</td>
<td>G</td>
<td>0.5551 (0.3288-0.937)</td>
<td>0.02614</td>
</tr>
<tr>
<td>5</td>
<td>rs3735035</td>
<td>PODXL</td>
<td>7q32</td>
<td>T</td>
<td>1.621 (1.042-2.522)</td>
<td>0.03126</td>
</tr>
<tr>
<td>6</td>
<td>rs10244329</td>
<td>LEP</td>
<td>7q33</td>
<td>T</td>
<td>2.062 (1.057-4.022)</td>
<td>0.03222</td>
</tr>
<tr>
<td>7</td>
<td>rs17632542</td>
<td>KLK3</td>
<td>19q13.33</td>
<td>C</td>
<td>0.313 (0.0938-1.043)</td>
<td>0.04647</td>
</tr>
<tr>
<td>8</td>
<td>rs3918256</td>
<td>MMP9</td>
<td>20q13.12</td>
<td>G</td>
<td>0.6428 (0.4128-1.001)</td>
<td>0.04959</td>
</tr>
</tbody>
</table>

Table 5.3 Statistically significant case-control SNP association between patients with aggressive prostate cancer and healthy controls

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene location</th>
<th>Tested allele</th>
<th>Odds Ratio (95% CI)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs17793693</td>
<td>PPAR-γ</td>
<td>3p25</td>
<td>A</td>
<td>4.534 (1.92-10.7)</td>
<td>0.000173</td>
</tr>
<tr>
<td>2</td>
<td>rs7931342</td>
<td>MYEOV</td>
<td>11q13</td>
<td>T</td>
<td>1.565 (1.206-2.032)</td>
<td>0.0007423</td>
</tr>
<tr>
<td>3</td>
<td>rs10896438</td>
<td>MYEOV</td>
<td>11q13.3</td>
<td>G</td>
<td>0.6673 (0.5141-0.8663)</td>
<td>0.002322</td>
</tr>
<tr>
<td>4</td>
<td>rs887391</td>
<td>SLC26A6</td>
<td>19q13</td>
<td>C</td>
<td>1.594 (1.148-2.213)</td>
<td>0.005094</td>
</tr>
<tr>
<td></td>
<td>SNP ID</td>
<td>Gene</td>
<td>Chromosome</td>
<td>Allele</td>
<td>OR (95% CI)</td>
<td>P-value</td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td>--------</td>
<td>------------</td>
<td>--------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>5</td>
<td>rs5945619</td>
<td>NUDT11</td>
<td>Xp11</td>
<td>C</td>
<td>0.5902 (0.4055-0.8592)</td>
<td>0.005749</td>
</tr>
<tr>
<td>6</td>
<td>rs17632542</td>
<td>KLK3</td>
<td>19q13.33</td>
<td>C</td>
<td>0.5005 (0.2969-0.8438)</td>
<td>0.008268</td>
</tr>
<tr>
<td>7</td>
<td>rs7920517</td>
<td>MSMB</td>
<td>10q11</td>
<td>G</td>
<td>0.7142 (0.5486-0.9298)</td>
<td>0.01227</td>
</tr>
<tr>
<td>8</td>
<td>rs11228565</td>
<td>MYEOV</td>
<td>11q13</td>
<td>A</td>
<td>0.6977 (0.5125-0.95)</td>
<td>0.02189</td>
</tr>
<tr>
<td>9</td>
<td>rs10244329</td>
<td>LEP</td>
<td>7q33</td>
<td>T</td>
<td>1.557 (1.06-2.286)</td>
<td>0.02344</td>
</tr>
<tr>
<td>10</td>
<td>rs6983561</td>
<td></td>
<td>8q24</td>
<td>C</td>
<td>0.5305 (0.298-0.9443)</td>
<td>0.02883</td>
</tr>
<tr>
<td>11</td>
<td>rs130067</td>
<td>CCHCR1</td>
<td>6p21.3</td>
<td>C</td>
<td>0.723 (0.533-0.9806)</td>
<td>0.03656</td>
</tr>
<tr>
<td>12</td>
<td>rs2727270</td>
<td>FADS2</td>
<td>11q13</td>
<td>T</td>
<td>0.6556 (0.4356-0.9867)</td>
<td>0.04184</td>
</tr>
<tr>
<td>13</td>
<td>rs12529</td>
<td>AKR1C3</td>
<td>10p15</td>
<td>G</td>
<td>0.7725 (0.5987-0.9967)</td>
<td>0.04685</td>
</tr>
<tr>
<td>14</td>
<td>rs2659122</td>
<td>KLK3</td>
<td>19q13.33</td>
<td>G</td>
<td>0.743 (0.5536-0.9972)</td>
<td>0.04748</td>
</tr>
</tbody>
</table>

**Tables 5.1- 5.3 colour legends risk association**

| Common SNPs between (Agg PCa vs Healthy Control) and (Agg PCa vs Non-Agg PCa) |
| Common SNPs between (Non-Agg PCa vs Healthy Control) and (Agg PCa vs Non-Agg PCa) |

### 5.1.4.2. Pathology, BMI and lifestyle

Owing to the established role of environmental or non-genetic aspects in the expression of genes (Labayen, Ruiz et al. 2011) in PCa (Gann 2002), we compared the variation in demographics and lifestyle factors such as age, alcohol consumption, smoking tobacco, and
levels of obesity among the patients recruited for our study. The disease association with BMI (based on the classification by World Health Organization (2000)) at the time of recruitment, tobacco smoking status, and alcohol consumption (at the time of recruitment in this study) are provided in Table 5.4, Table 5.5, and Table 5.6 respectively. In these tables, the controls are compared to non-aggressive and aggressive PCa patients.

Table 5.4 The association between BMI and risk and/or aggressiveness of prostate cancer

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Pathology</th>
<th>N' (N &amp; UW)</th>
<th>(OW &amp; O)</th>
<th>Total</th>
<th>Median (75th Percentile)</th>
<th>Mann-Whitney U test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive vs Control</td>
<td>Aggressive</td>
<td>36</td>
<td>161</td>
<td>197</td>
<td>27.00 (30.50)</td>
<td>31166.00</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>122</td>
<td>247</td>
<td>369</td>
<td>26.36 (29.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggressive vs Non-Aggressive</td>
<td>Aggressive</td>
<td>36</td>
<td>161</td>
<td>197</td>
<td>27.00 (30.50)</td>
<td>5450.50</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>Non-Aggressive</td>
<td>8</td>
<td>49</td>
<td>57</td>
<td>27.00 (29.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Aggressive vs Control</td>
<td>Non-Aggressive</td>
<td>8</td>
<td>49</td>
<td>57</td>
<td>27.00 (29.00)</td>
<td>9115.00</td>
<td>0.105</td>
</tr>
</tbody>
</table>

Control 122 247 369 26.36 (29.12)

Table legends: N= men with normal weight (18.50 kg/m² - 24.99 kg/m²); UW= men who were under-weight (<18.50 kg/m²); OW= men who were over-weight (25.00 kg/m² - 29.99 kg/m²); O= men who were obese (≥30.00 kg/m²); N’= number of men
Table 5.5 Association between tobacco smoking status and risk and/or aggressiveness of prostate cancer

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Pathology</th>
<th>N¹</th>
<th>Percentage (ever smokers)</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
<td>54.82%</td>
</tr>
<tr>
<td>Aggressive vs Control</td>
<td>Aggressive</td>
<td>89</td>
<td>108</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>225</td>
<td>144</td>
<td>369</td>
<td>39.02%</td>
</tr>
<tr>
<td>Aggressive vs Non-</td>
<td>Aggressive</td>
<td>89</td>
<td>108</td>
<td>197</td>
<td>54.82%</td>
</tr>
<tr>
<td>Non-Aggressive</td>
<td>Non-Aggressive</td>
<td>23</td>
<td>34</td>
<td>57</td>
<td>59.64%</td>
</tr>
<tr>
<td>Non-Aggressive vs</td>
<td>Non-Aggressive</td>
<td>23</td>
<td>34</td>
<td>57</td>
<td>59.64%</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>225</td>
<td>144</td>
<td>369</td>
<td>39.02%</td>
</tr>
</tbody>
</table>

Table legends: N¹= number; 95% CI= 95% confidence interval, No= never smokers, Yes= ever smoker

Table 5.6 Comparison of alcohol consumption and risk and/or aggressiveness of prostate cancer

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Pathology</th>
<th>N¹</th>
<th>Percentage (alcohol consumers)</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No (1)</td>
<td>Yes (0)</td>
<td>Total</td>
<td>59.89%</td>
</tr>
<tr>
<td>Aggressive vs Control</td>
<td>Aggressive</td>
<td>79</td>
<td>118</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>54</td>
<td>315</td>
<td>369</td>
<td>85.36%</td>
</tr>
<tr>
<td>Aggressive vs</td>
<td>Aggressive</td>
<td>79</td>
<td>118</td>
<td>197</td>
<td>59.89%</td>
</tr>
<tr>
<td>Non-Aggressive</td>
<td>Non-Aggressive</td>
<td>25</td>
<td>32</td>
<td>57</td>
<td>56.14%</td>
</tr>
<tr>
<td>Non-Aggressive vs</td>
<td>Non-Aggressive</td>
<td>25</td>
<td>32</td>
<td>57</td>
<td>56.14%</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>54</td>
<td>315</td>
<td>369</td>
<td>85.36%</td>
</tr>
</tbody>
</table>

Table legends: N¹= number; 95% CI= 95% confidence interval
5.1.4.3. **Correction for the effect of covariates to identify the genic risk of aggressive PCa**

The statistically significant results obtained using logistic models after adjustments for multiple covariates are presented in Table 5.7 and Table 5.8 (Aggressive PCa vs non-aggressive PCa and aggressive PCa vs healthy control are shown in Table 5.7 and Table 5.8 respectively). The data obtained for non-aggressive PCa vs healthy control was not statistically significant and will not be further addressed in this paper.

The results obtained after the interaction analysis can also be broadly classified into two categories- expected, and unique to the New Zealand population studied. The New Zealand aspect of gene x environment interaction was much better evident after analyzing these results.

Certain SNPs present near genes such as *MYEOV* - a putative oncogene (Table 5.8); *KLK3* and *MSMB* (both, Table 5.8)– involved in the androgen metabolism pathway; *MMP9* (Table 5.7)- involved with the inflammatory pathway; and *MLH1* (Table 5.8), which has a role in DNA mismatch repair were identified as statistically significant in our study as expected. However, we had some novel findings as well.

A SNP present near the gene *LEP* (Table 5.7 and Table 5.8), associated with obesity is identified as statistically significant risk for both aggressive and non-aggressive PCa; and the SNP present near the gene *SEP15* (*Selenoprotein 15KDa*) - involved with quality control of protein folding in the endoplasmic reticulum (ER) may reveal new knowledge about the changes of risk for aggressive PCa with local environmental conditions and its effects on the New Zealand Caucasian men studied here.
Table 5.7 Logistic model of SNPs risk for aggressive prostate cancer vs non-aggressive prostate cancer after correcting for interaction with multiple covariates (BMI, tobacco smoking, and alcohol consumption)

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene location</th>
<th>Tested allele</th>
<th>Odds Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs632148</td>
<td>SRD5A2</td>
<td>2p23.1</td>
<td>C</td>
<td>2.144</td>
<td>0.01196</td>
</tr>
<tr>
<td>2</td>
<td>rs10244329</td>
<td>LEP</td>
<td>7q33</td>
<td>T</td>
<td>2.395</td>
<td>0.02893</td>
</tr>
<tr>
<td>3</td>
<td>rs3918256</td>
<td>MMP9</td>
<td>20q13.12</td>
<td>A</td>
<td>1.873</td>
<td>0.03109</td>
</tr>
<tr>
<td>4</td>
<td>rs3735035</td>
<td>PODXL</td>
<td>7q32</td>
<td>T</td>
<td>1.625</td>
<td>0.03561</td>
</tr>
</tbody>
</table>

Table 5.8 Logistic model of SNPs risk for aggressive prostate cancer vs healthy controls after correcting for interaction with multiple covariates (BMI, tobacco smoking, and alcohol consumption)

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene location</th>
<th>Tested allele</th>
<th>Odds Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs10896438</td>
<td>MYEOV</td>
<td>11q13</td>
<td>T</td>
<td>1.699</td>
<td>0.000421</td>
</tr>
<tr>
<td>2</td>
<td>rs7931342</td>
<td>MYEOV</td>
<td>11q13</td>
<td>T</td>
<td>1.705</td>
<td>0.000423</td>
</tr>
<tr>
<td>3</td>
<td>rs7920517</td>
<td>MSMB</td>
<td>10q11</td>
<td>A</td>
<td>1.590</td>
<td>0.001734</td>
</tr>
<tr>
<td>4</td>
<td>rs5845</td>
<td>SEP15</td>
<td>1p22.3</td>
<td>C</td>
<td>1.845</td>
<td>0.01333</td>
</tr>
<tr>
<td>5</td>
<td>rs2659122</td>
<td>KLK3</td>
<td>19q33.33</td>
<td>A</td>
<td>1.702</td>
<td>0.01516</td>
</tr>
<tr>
<td>6</td>
<td>rs1799977</td>
<td>MLHI</td>
<td>3p21</td>
<td>G</td>
<td>1.53</td>
<td>0.0202</td>
</tr>
<tr>
<td>7</td>
<td>rs10244329</td>
<td>LEP</td>
<td>7q33</td>
<td>T</td>
<td>1.639</td>
<td>0.03245</td>
</tr>
<tr>
<td>8</td>
<td>rs12529</td>
<td>AKR1C3</td>
<td>10p15</td>
<td>C</td>
<td>1.318</td>
<td>0.04006</td>
</tr>
<tr>
<td>9</td>
<td>rs887391</td>
<td>SLC26A6</td>
<td>19q13</td>
<td>C</td>
<td>1.639</td>
<td>0.04449</td>
</tr>
</tbody>
</table>

Tables 5.7- 5.8 colour legend showing risk association

|                  | Common SNPs between (Agg PCa vs Healthy Control) and (Agg PCa vs Non-Agg PCa) classifications |

5.1.5. Discussion

It is well established that there are three major risk factors for PCa, namely, advancing age (Vaidyanathan, Karunasinghe et al. 2016), ethnicity, and familial history (Gann 2002). Various studies have indicated that alterations in genetic and epigenetic make-up are predominantly the basis for the development of various malignancies (Delpu, Hanoun et al. 2011; Lund, Narva et al. 2012; Goh, Saunders et al. 2013). In this study, SNP genotype data
was used to identify risk association with aggressive PCa while the effect of non-genetic or environmental factors was also considered.

5.1.5.1. Effect of environment and SNP genotype with risk of prostate cancer

Parts of the panel of SNPs used in this analysis have previously been considered to assess PCa risk (Supplementary table 1). However, the approach here is to assess their significance to aggressive PCa also. Out of the 97 SNPs studied, 4 SNPs were significantly associated with non-aggressive PCa when compared with healthy controls, and 8 SNPs were identified to be significantly associated with aggressive PCa compared to non-aggressive PCa. Also, 14 SNPs were identified to be significantly associated with aggressive PCa when compared with healthy controls (Tables 5.1- 5.3 respectively). The majority of the SNP associations lost significance after correcting for multiple testing using the BONF, an overly conservative approach, and the BH-FDR, a less conservative correction that tolerates more false positives.

Large scale GWAS have previously shown direct SNP associations with aggressive PCa. However, SNP interactions with demographic and lifestyle factors could also add to the allelic effect producing a modified risk of a disease. These SNPs could be indicating a unique situation for New Zealand men- an example of isolated men of European origin, with PCa. Of particular interest in the New Zealand context are, aspects such as obesity (2015), higher intake of red meat and dairy products, and possible deficiency in selenium intake. All of these could impact on genetic mechanisms in ways that may lead to a higher risk of aggressive PCa.

The three broad classifications under which the data were analyzed are SNP associations between patients with aggressive PCa and healthy controls, between patients with non-aggressive PCa and healthy controls and between patients with aggressive and non-aggressive PCa. The results for these groups are discussed below.
5.1.5.1.a. SNP genotype analysis of non-aggressive prostate cancer vs healthy controls

Of the 4 SNPs identified as significant risk for non-aggressive PCa vs healthy controls, 3 SNPs were identified to be commonly associated to risk of the progression of this disease (aggressive PCa vs non-aggressive PCa). These SNPs include rs2292884 in chromosomal region 2q37 near the gene MLPH (Melanophilin), rs3735035 present in chromosomal region 7q32 in the gene PODXL (Podocalyxin-like), rs10086908 present in chromosomal region 8q24 and rs11536889 present in chromosomal region 9q33 near the gene TLR4.

The TLR4 gene is responsible for activating innate immunity in humans (Nahid, Satoh et al. 2011; Zhao, Zhang et al. 2014). TRL4 is one of 13 transmembrane receptors found in the troll-like receptors family which plays an important role in chronic infection and inflammation pathways, in turn controlling the incidence of development of cancer (Rakoff-Nahoum and Medzhitov 2009; Zhao, Zhang et al. 2014). The expression of the TLR4 protein in PCa is already well established (Zhao, Zhang et al. 2014).

SNP rs2292884 close to chromosomal region 2q37 near the gene MLPH, was also identified to be statistically significantly associated with non-aggressive PCa compared to controls. Very little is known about the possible functional impact of MLPH in carcinogenesis and/or tumour progression (Bu, Narisu et al. 2016). Dysregulation of the protein MLPH has recently been found in several types of tumours in lung cancer, meningiomas, and of breast cancers (Fevre-Montange, Champier et al. 2009; Pio, Blanco et al. 2010; Thakkar, Raj et al. 2010). A recent study found an association of expression of the gene MLPH with nearby SNPs in prostate tissue (Penney, Sinnott et al. 2015). In non-small cell lung cancer, MLPH mRNA was identified as a target of differentially expressed miRNAs (Molina-Pinelo, Gutierrez et al. 2014). Interestingly, the protein MLPH was also found to be significantly overexpressed in estrogen receptor (ER) positive breast cancer, suggesting a regulation of this protein by estrogen hormones (Thakkar, Raj et al. 2010).
SNP rs3735035 present in chromosomal region 7q32 in the gene *PODXL* was also identified to be significantly associated with non-aggressive PCa (before the use of BONF) in our study. *PODXL* is a cell-adhesion glycoprotein and stem cell marker that has been associated with an aggressive tumour phenotype and poor prognosis in several forms of cancer (Larsson, Johansson et al. 2011; Boman, Larsson et al. 2013). Interestingly, the first report of *PODXL* expression in malignant cells was its description as a stem cell marker in testicular cancer (Schopperle, Kershaw et al. 2003) and it has been previously identified as a candidate biomarker for PCa aggressiveness (Casey, Neville et al. 2006). Since rs3735035 present in this gene came up as statistically significantly associated with non-aggressive PCa (before the use of BONF) in our population, we feel that this SNP may play a role from the point of disease initiation.

SNP rs4965373 present in chromosomal region 15q26.3 in the gene *SEPS1* (*Selenoprotein S*) was also identified to be statistically significantly associated with non-aggressive PCa (before the use of BONF) in our population. The human gene *SEPS1*, encodes selenoprotein S which participates in the retro-translocation of misfolded proteins from the ER to the cytosol for their degradation (Ye, Shibata et al. 2004). This ER membrane protein functions in stress responses to prevent the deleterious consequences of accumulation of misfolded proteins, that has been linked to immune and inflammatory processes (Pahl and Baeuerle 1997). The *SEPS1* gene was first suggested to be related to the stress response process including immune and inflammatory processes through the study of Curran *et al.*, in 2005(Curran, Jowett et al. 2005). In addition, several diseases, including inflammatory disorders(Martinez, Santiago et al. 2008) such as insulin-dependent diabetes mellitus (Field, Tobias et al. 1994), and gerontological disorders such as Alzheimer’s disease (Blacker, Bertram et al. 2003), have been shown to be linked to this gene region. The A allele of this rs4965373 SNP has previously shown positive associations with serum Selenium levels in a
healthy male population from New Zealand (Lynnette, Nishi et al. 2012). Comparing the current findings with those of Ferguson et al. (2012) indicates a possibility for retention of excess selenium as a cause of non-aggressive PCa risk. These direct us to the relation between functional defects, inflammation, immunity, and its depletion due to ageing as responsible for initiation of diseases such as PCa.

5.1.5.1.b. **SNP genotype analysis of aggressive prostate cancer vs non-aggressive prostate cancer**

Non-aggressive PCa can be considered as an intermediary between healthy and aggressive PCa states. We identified a number of SNPs to be common between the categories of SNP association between aggressive PCa and healthy controls and non-aggressive PCa and healthy controls.

Among these common SNPs are those in genes that have been identified to play roles in immunity (rs11536889) (Beutler 2009), development and progression of PCa (rs2292884 (Bu, Narisu et al. 2016), and rs3735035 (Liu, Cheng et al.)).

The SNPs common to be associated with aggressive PCa compared to controls and the SNPs identified to be associated with the progression of this disease (aggressive PCa vs non-aggressive PCa) are rs10244329 in 7q33.3 chromosomal region near the gene LEP, rs887391 and rs17632542, both present in chromosomal region 19q13. Of the SNPs that are common to be associated to non-aggressive PCa compared to controls and the SNPs identified to be associated with the progression of this disease (aggressive PCa vs non-aggressive PCa), there are SNPs in genes that have been identified to play roles in the expression of PSA (rs17632542) (Kote-Jarai, Amin Al Olama et al. 2011; Sullivan, Kopp et al. 2015), development and progression of PCa (rs887391) (Hsu, Sun et al. 2009), and morbid obesity (Farooqi and O'Rahilly 2008; Beckers, Zegers et al. 2009).
SNPs rs632148 in the gene \textit{SRD5A2} (\textit{Steroid 5\textalpha-reductase type 2}) on chromosomal region 2p23, and rs3918256 in the gene \textit{MMP9} have also been associated with aggressive PCa when compared to non-aggressive PCa patients. The SNPs rs632148 and rs3918256 have been previously reported by groups working on various aspects related to and causing PCa in Caucasian populations and not restricted to studies pertaining to quality of sperms (rs632148) (Zhao, Wu et al. 2012), and also contribute to the invasive and metastatic properties of malignant tumours, including those of the prostate (rs3918256) (Aalinkeel, Nair et al. 2011). These SNPs are of considerable importance in understanding the overall progression of PCa to aggressive stage (Orozco, Goh et al. 2013).

\textbf{5.1.5.1.c. SNP genotype analysis of aggressive prostate cancer vs healthy controls}

A link between SNPs in the 11q13 region and PCa has been previously identified by various GWAS (Eeles, Kote-Jarai et al. 2008; Thomas, Jacobs et al. 2008; Chung, Ciampa et al. 2011; Lange, Salinas et al. 2012). Three SNPs- rs10896438, rs7931342, rs11228565, present in \textit{MYEOV} and one SNP- rs2727270 present in \textit{FADS2} are identified to be statistically significant before the use of BONF when comparing the controls to those with aggressive PCa. Frequent rearrangements are observed in human cancers in the area 11q13 (Song, Koessler et al. 2008). Four out of fourteen SNPs identified to have a statistically significant risk association with aggressive PCa, were identified in one location. This region houses a number of candidate oncogenes, and the amplification of this region is highly heterogeneous leading to breakpoints in and/or near the location (Jarmuz-Szymczak, Pelinska et al. 2013; Szyfter, Wierzbicka et al. 2016). \textit{MYEOV} has been shown to be frequently amplified in tumours not only of the breast but also of the oesophagus (Brecht, Steenvoorden et al. 2007) and head and neck squamous cell carcinoma (Szyfter, Wierzbicka et al. 2016).
Of the 3 SNPs present in 19q13 region, 2 (rs2659122 and rs17632542) are located in the gene region *KLK3*, which codes for PSA- a glycoprotein enzyme, secreted by epithelial cells of the prostate gland. Men with serum PSA levels of more than 20ng/ml are categorized as having aggressive PCa (Thompson, Thrasher et al. 2007). Therefore, it is not surprising that a number of SNPs in the gene *KLK3* were identified as statistically significant when comparing men with aggressive PCa and controls. This is similar to the findings discussed in other GWAS for aggressive PCa (Eeles, Kote-Jarai et al. 2008). The third SNP in this region, rs887391, present in gene *SLC26A6* (*Solute carrier family 26 member 6*)- a fusion gene which is crucial for the development and the progression of various human cancers (Lambros, Wilkerson et al. 2011), is about 10Mb centromeric to the *KLK3* gene, where the SNP rs2735839, near 3’ end is found (Eeles, Kote-Jarai et al. 2008; Hsu, Sun et al. 2009).

The chromosomal region, 8q24 has been of considerable interest in terms of development and epidemiology of cancer (Park, Chang et al. 2008). Ahmadiyeh *et al.*, (2010) suggested that numerous, non-dependent polymorphic variants present in the chromosome location 8q24 may produce certain biological mechanisms that contribute to disease, or, alternatively, the 8q24 regions may cumulatively influence the regulation of adjacent genes (cis-regulation) or genes on other chromosomes (trans-regulation) (Ahmadiyeh, Pomerantz et al. 2010). The SNP rs6983561 identified to be significantly associated with aggressive PCa compared to controls among our population before the use of BONF, however, is commonly found in the Asian population, rather than the Caucasian population (Suzuki, Liu et al. 2011).

10q11 (Camp, Farnham et al. 2009; FitzGerald, Kwon et al. 2009) and Xp11 (Camp, Farnham et al. 2009; FitzGerald, Kwon et al. 2009; Lange, Salinas et al. 2012) are two other chromosomal regions that are identified to have a number of SNPs associated with PCa in Caucasian populations. We identified one SNP, in each of the two aforementioned locations to be significantly associated with aggressive PCa compared to controls before the use of
BONF. These are rs7920517 present near the gene MSMB, and rs5945619 present in the gene NUDT11 (Nucleoside Diphosphate-linked Moiety X Motif 11) respectively. These SNPs knit a very tight story and are of considerable interest.

MSMB is one of just three predominant proteins, along with PSA and prostatic acid phosphatase secreted by a normal human prostate gland (Lilja and Abrahamsson 1988). The PSA level in men is usually measured to estimate the health of a prostate gland, and MSMB, owing to its varied expression levels in cancer cells compared to the prostate epithelial cells in the benign stage, is implicated as a potential PCa biomarker (Maeda, Kamada et al. 1994; Sjoblom, Saramaki et al. 2016). The protein MSMB is also known to modulate immunoglobulin levels in hypoxic tumour cores (Karunasinghe N 2014).

The SNP rs5945619 (Fitzgerald, Kwon et al. 2009; Lange, Salinas et al. 2012) present in the Xp11 region also has been identified to be significantly (before BONF were implemented) associated with aggressive PCa to controls. This SNP is present near the paralogous human gene NUDT11, predominantly expressed in the testes, and may play a role in signal transduction (Hidaka, Caffrey et al. 2002; Hua, Hidaka et al. 2003). A number of GWAS have also suggested that the susceptibility locus at NUDT11 may have involvement with risk of PCa (Eeles, Kote-Jarai et al. 2008; Gudmundsson, Sulem et al. 2008; Camp, Farnham et al. 2009; Fitzgerald, Kwon et al. 2009; Lu, Sun et al. 2009).

The SNP rs17793693 present in the PPAR-γ gene was also identified to be significantly associated with aggressive PCa in our population. The significance, however, was lost after the BONF was implemented. PPAR-γ ligands induce growth arrest in cells through apoptosis, in both macrophages (Chinetti, Griglio et al. 1998) and endothelial cells (Bishop-Bailey and Hla 1999). Among other roles, the PPAR-γ protein has also been identified to be having a role in immunity (Braissant, Foulfelle et al. 1996; Segawa, Yoshimura et al. 2002). The PPAR-γ gene, present in the 3p25 region of the human genome has been suggested to have an
association with PCa by various researchers (Segawa, Yoshimura et al. 2002; Eeles, Olama et al. 2013). PPARs (comprising of PPAR-α, PPAR-β and PPAR-γ genes) are members of the nuclear hormone receptor superfamily of ligand-activated transcription factors including receptors for steroid hormones, Vitamin D3, retinoid and thyroid hormones (Issemann and Green 1990; Segawa, Yoshimura et al. 2002). The PPAR-γ gene has been identified to be mainly present in adipose tissues, wherein it plays a vital role in the pre-adipocytes to adipocytes (Segawa, Yoshimura et al. 2002). The PPAR-γ gene has also been identified to be involved in the pathology and progression of various diseases including obesity and diabetes mellitus (Derosa and Maffioli 2012). Diabetes mellitus has previously been associated with PCa, such that patients with diabetes have been suggested to be at a lower risk of PCa (Giovannucci, Harlan et al. 2010). Certain metabolic factors including reduced testosterone levels may affect blood glucose level (Ma, Li et al. 2008; Giovannucci, Harlan et al. 2010).

SNPs rs130067 present in 6p21.3 chromosomal region near the gene CCHCR1 (Coiled-coil alpha-helical rod protein1), rs10244329 present in 7q33.3 chromosomal region near the gene LEP and rs12529 present in chromosomal region 10p15 in the gene AKR1C3 (Aldo-keto reductase family 1 member C3) have also been identified to be statistically significantly associated with aggressive PCa in this study. The AKR1C3 rs12529 G allele has been previously associated with PCa risk when interacting with age and lifestyle habits including tobacco smoking (Karunasinghe, Lange et al. 2013). It could be that although the G allele of this SNP is producing a risk for PCa, it is the C allele that produces aggressiveness of the disease. A study by Yu et al., (2013) (Yu, Huang et al. 2013) has shown that this C allele is responsible for prostate cancer- specific mortality among those receiving androgen deprivation therapy. The same allele has also shown retention of cancer-specific symptoms, if managed only with non-androgen deprivation treatment methods (Karunasinghe, Zhu et al. 2016). The SNPs rs130067, rs10244329, and rs12529 have been previously reported by
groups working on various diseases related to and causing PCa in Caucasian populations such as rheumatoid arthritis (Orozco, Goh et al. 2013), obesity (Labayen, Ruiz et al. 2011), and sex hormone metabolic pathway (Adeniji, Chen et al. 2013; Karunasinghe, Lange et al. 2013; Huang, Chien et al. 2015) respectively. The rs12529 SNP is also related to lung cancer (Lan, Mumford et al. 2004) and bladder cancer (Figueroa, Malats et al. 2008).

5.1.5.2. Age and risk of prostate cancer

Age is a major risk factor for PCa (Vaidyanathan, Karunasinghe et al. 2016) (Karunasinghe, Han et al. 2013). However, in the data presented in the current section we did not consider the role of ageing, as it is the only external factor which cannot be reversed, but the other aspects such as high (or, low) BMI, smoking tobacco and alcohol consumption habits can be altered by individuals. Moreover, we wanted to see the effect of gene x environment aspects prevalent in our local population in the expression and progression of PCa. The section 5.4 is dedicated to analyse the effect of age as a risk factor for aggressive PCa.

5.1.5.3. BMI, smoking tobacco, and alcohol consumption at recruitment and risk of prostate cancer

Our previous research showed that BMI is not statistically different in age matched controls and those with benign urology disease (Karunasinghe, Han et al. 2012). However, with the current stratification BMI was shown to be significantly higher (27.00kg/m²) among those with aggressive PCa compared to controls (26.36kg/m²). The mean BMI at recruitment of all our groups (cases as well as healthy controls) was calculated to be in the category of overweight for adults. However, Vidal et al. (2014) (Vidal, Howard et al. 2014) and Haque et al (2014) (Haque, Van Den Eeden et al. 2014) discuss that men with aggressive PCa have a higher risk of mortality due to the disease if they are overweight or obese.
Smoking tobacco has been identified as a risk factor for PCa (Huncharek, Haddock et al. 2010), and our results support smoking tobacco being a major risk factor in individuals developing aggressive as well as non-aggressive PCa. Our results suggest that more than half of our patient cohort identified themselves as ever smokers, and this is very similar to the results discussed by Huncharek et al. (2010) (Huncharek, Haddock et al. 2010) and Braithwaite et al. (2012) (Braithwaite, Boffetta et al. 2012). In another study carried out by Pantarotto et al. (2007) (Pantarotto, Malone et al. 2007), previous and present tobacco smokers have been associated with a greater risk of PCa metastasizing. Tobacco smoking, has previously been reported as a risk factor for PCa as a whole, compared with age and ethnicity matched healthy controls by our group (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2012). A systematic review in 2014 indicated that tobacco smoking is a risk factor for fatal prostate cancer but not a risk factor for prostate cancer incidence (Islami, Moreira et al. 2014). Adding to the published work by Islami et al. (2014), Schlatmann and Blanker (2015) indicated that the mortality risk due to tobacco use was higher in the pre-PSA era and not in the post-PSA era (Islami, Moreira et al. 2014; Schlatmann and Blanker 2015). Tobacco use being a risk factor for both aggressive and non-aggressive PCa in our New Zealand cohort is concerning as it indicates a scenario similar to that of the pre-PSA era from other western countries. The possibility of former tobacco smokers changing their lifestyle after being identified as having PCa, or being in a high-risk population, has also been reported by Blanchard et al. (2003). However, unlike the alcohol consumption changes observed in our study cohorts, the frequency of men ever exposed to tobacco smoking in both aggressive (54.8%) and non-aggressive (51.06%) groups has not declined (53.1% in 2012) (Karunasinghe, Han et al. 2013).

Alcohol consumption, in spite of being well-documented as a potential cause of cancer (Bagnardi, Rota et al. 2015; Connor 2016), is yet to be proven as a risk factor for PCa (Baan,
Straif et al. 2007; Zuccolo, Lewis et al. 2013). Unlike the results of Zuccolo et al. (2013), who have shown a moderate risk of aggressiveness with alcohol consumption (Zuccolo, Lewis et al. 2013), our study has found inverse significance related to alcohol consumption when comparing data obtained from groups of aggressive PCa vs controls as well as non-aggressive PCa vs controls. Additionally, previous publications from our group with the initial 264 PCa patients recruited within one year of PCa diagnosis, has recorded that 72% of patients were alcohol consumers (Karunasinghe, Han et al. 2012). The current decline in alcohol consumption rate to 59% in patients with aggressive PCa and 56% in patients with non-aggressive PCa could well be due to PCa survivors changing their behavior and consuming less alcohol.

### 5.1.5.4. Correction for gene x environment interaction and risk of prostate cancer

Knowledge of gene–environment interaction is important for risk prediction and the identification of certain high-risk populations to inform public health strategies for targeted prevention (Rudolph, Chang-Claude et al. 2016). We associated the environmental factors with the genotypes of the men in our study to identify the risk alleles for specific modifiable factors such as BMI, smoking tobacco and alcohol consumption. Since these factors play an important role in the risk association of PCa and yet can be controlled by individuals, they therefore are of much importance to understand and limit this disease.

#### 5.1.5.4.a. Genotype and the outcome of non-aggressive prostate cancer (vs healthy controls)

Many malignancies have been linked to specific environmental exposures (Mullins and Loeb 2012). Several environmental and occupational factors such as farming and use of pesticides, exposure to sunlight/ultraviolet radiation, as well as trace minerals which are commonly used in tyre and battery manufacturing have been studied for an association with
PCa risk (Mullins and Loeb 2012). The factors, however, that initiate PCa, unfortunately, still remain to be identified as an essential prelude to strategies designed to reduce disease occurrence (Grover and Martin 2002).

Figure 5.2 Individual aggressive PCa risk association with SNPs, environmental factors as well as in combination of both compared to non-aggressive disease.
In our study, we have certain indicators in terms of the external or environmental factors such as tobacco smoking, alcohol consumption, or high BMI associated with the initiation of this disease, i.e. with non-aggressive PCa. However, when the data was corrected for these aspects using PLINK and re-analyzed, not one of the SNPs were identified as significantly associated with the disease. This may be because mammalian body is designed to fight unfavorable conditions (Elsberg 1899) and it is the external factors which influence the initiation of non-aggressive PCa.

5.1.5.4.b. Genotype and the outcome of aggressive prostate cancer (vs non-aggressive prostate cancer)

We corrected for the effect of three prominent risk factors for the prognosis and progression of PCa- BMI, tobacco smoking and alcohol consumption as risk factors for aggressive PCa (compared to non-aggressive PCa) using PLINK and then analyzed the data to identify a pure genic risk for the disease and are mentioned in Table 5.7. A holistic approach of these external factors was taken into consideration while looking into the gene-environment interaction in our study.

Only few SNPs were identified to have statistically significant risk for aggressive PCa risk compared to that of non-aggressive disease (before the use of BONF). They have been illustrated in Figure 5.2.

SNPs rs632148, rs10086908, rs887391, and rs16987929 were identified to be common both with genotypic risk as well as interacting with environmental risk. Common SNPs between these states indicate the crucial role of gene x environment interactions in both initiation and the progression of this disease.
5.1.5.4.c. **Genotype and the outcome of aggressive prostate cancer (vs healthy controls)**

We continued to analyze the effect of the three aforementioned risk factors on the genotype in our population for the prognosis of PCa by statistically correcting it using PLINK. The same holistic approach as previously mentioned, in terms of the external factors were taken into consideration while looking into the gene-environment interaction in our study.

A higher number of SNPs were identified to have statistically significantly increased interactions with external factors in producing aggressive PCa compared to that of healthy controls (before the use of BONF) than the SNPs which potentially had a genic effect on the disease. They have been summarized in Figure 5.3.

Of the 9 SNPs that came up significantly associated with aggressive PCa after correcting for external factors in our population, only 2 SNPs were identified not to be repeats from the results already discussed in section 4.1.3. The higher number of SNPs identified in the gene x environment interaction category compared to in this category, provides statistical proof that gene x environment interactions play a crucial role in the progression of this disease. Of the SNPs that are identified in the genes, the expression of which may be independent of the environmental, or external factors (Table 5.8), were rs5845 and rs1799977.

The SNP rs5845 present in the gene *SEP15* encoding for selenoprotein containing selenocysteine is involved in the quality control of protein folding (Korotkov, Kumaraswamy et al. 2001). The A allele of this SNP has a unique effect on PCa risk even compared to benign urology disease (Karunasinghe, Han et al. 2012). The SNP rs1799977 is present in the gene *MLH1*. The gene *MLH1* plays a major role in mismatch repair (Pal, Permuth-Wey et al. 2008). Although we did not analyse SNPs in all the DNA repair genes, but still owing to the fact that we considered tobacco smoking- an important risk of DNA damage (Lockett, Hall et al. 2006) and found statistically significant association with risk of aggressive PCa among our
cohort, we believe that this is a good co-relation with our findings and the results herewith discussed. The SNP rs10896469 present near MYEOV, a putative oncogene, as previously mentioned to be frequently amplified in a number of tumours (Brecht, Steenvoorde et al. 2007; Colloca and Venturino 2011; Szyfter, Wierzbicka et al. 2016).

Interestingly, certain SNPs showing significant associations as risk for aggressive PCa did not show significant associations when interacting with environmental factors. In contrast, the SNPs that showed the strongest evidence for interactions with environmental factors did not show significant in SNP-only effect association with the risk of the disease.

This finding reveals that in a study like ours, it is very important to not only consider the SNPs identified by a GWAS, but also consider genotype (additive or dominant) by environmental (local to the population studied) interaction effects on aggressive PCa in addition to established associations. If this is avoided, such variants may be ignored. Therefore, the list of SNPs to be studied and eventually followed up for replication or confirmation changes with the genotype x environmental interaction effects should be extensive and not limited to just the ones reported in certain populations ignoring the local factors which may be pivotal for the final expression of genes.

We would also like to point out that, SNPs with statistically significant genotype x environmental interaction did not necessarily have a significant (or even suggestive) association with aggressive PCa.

Thus, limiting interaction tests to markers with significant main SNP effects would likely find different results, and will aid in identifying a much better localized cause of diseases, as has also been mentioned by Arya et al., (2009) (Arya, Hare et al. 2009). Since the genetics of aggressive PCa is still largely unknown, we cannot comment whether this approach would lose power or would protect against false positives because the underlying genetic
architecture of aggressive PCa is largely unknown and more research needs to be done in this field of study, but we believe this is the right direction to move forward.

Figure 5.3 Individual aggressive PCa risk association with SNPs, environmental factors as well as in combination of both compared to controls.
5.1.6. Conclusions

There is an urgent need to develop a biomarker for PCa with high sensitivity, and specificity to an individual’s risk of developing the aggressive form of the disease (Vesprini, Liu et al. 2013). A considerable number of patients with aggressive PCa are being under-treated (Cooperberg, Broering et al. 2010). The results obtained are suggestive of certain genes being associated with aggressive PCa. The data generated, suggest that SNP genotyping as a screening tool, along with other prevalent diagnostic tools, has the potential to help identify men heading towards lethal aggressive PCa at an earlier stage.

The fact that a number of SNPs in various genes were identified to be associated with a risk of PCa calls for detailed work in this aspect for risk of this disease. We observed that the SNPs vulnerable to environmental conditions discussed here were not identified to be statistically associated with the initiation of the disease (perhaps due to the small sample size). However, the gene x environmental effect was identified to be more prominent with the progression of the disease. It is also worthy of mention that although we don’t believe that the Caucasian population of New Zealand is not genetically distinct, but the various external factors described here, such as increasing BMI (2015), and wide-spread tobacco smoking in New Zealand (2015), and deficiency of trace elements such as selenium in the New Zealand soil (Hewitt and Dymond 2013) may be playing a much more important role in the expression and progression of PCa in our population than they have been previously credited for and need to be further looked into. These factors may be very crucial for the progression of the disease and our gene x environment logic is strengthened by the findings in our unique population.

We identified that the SNPs rs17793693 in the gene PPAR-γ and rs2727270 in the gene FADS2 were risk for aggressive PCa before correction for environmental factors was carried out However, after correcting for multiple factors, these SNPs did not figure as a risk for
aggressive PCa. This may be indicative of yet another health concern to the higher percentage of individuals with high BMI in our population (2015). We also identified the SNP rs4965373 in the gene SEPS1 to be significantly associated with risk for non-aggressive PCa (vs healthy controls), but, again, when corrected for, did not come up as a risk association for the disease. Interestingly, though, we identified another SNP rs5845 in another selenoprotein gene- SEP15 when we corrected the SNP association for aggressive PCa (vs healthy controls) for multiple factors. This is very unique to the environmental conditions in New Zealand, where a selenoprotein SNP may trigger the start of the disease, but is eventually controlled due to the lack of selenium in the New Zealand soil, and is effectively checked in the eventual progress of the disease, unlike if it is the case when proper intake of selenium is not practiced. However, further works need to be done with these regards to further our findings.

These various associations and links leading to aggressive and non-aggressive PCa are better explained in the pictorial conclusion (Figure 5.4).

![Figure 5.4 Pictorial conclusion illustrating the various associations and links leading to aggressive and non-aggressive PCa.](image)

**Figure legends** - Non-Agg: Non-aggressive PCa patients; Agg: Aggressive PCa patients; Con: Healthy Controls
Supplementary Materials (Available online):

Supplementary Table 1: Selected SNPs for present study.

Supplementary Table 2: Gene x environment association analysis of SNPs as risk for PCa

Supplementary Table 3: SNP association with risk of PCa after correcting for environmental factors
5.2. Section summary and preamble to next section

It was interesting to identify the involvement of gene x environment interactions playing such a vital role in case of risk for aggressive PCa, especially with the strong association to the prevalent conditions and lifestyle factors in and among the population of New Zealand. There have been several studies prior to us, and shall be more to follow, but our broad and holistic approach was one of a kind. We took a clue from these findings, and since ageing is one of the major risk factors for risk of PCa, we then wanted to statistically analyse the effect of ageing by adjusting it statistically in our case-control study.

Sections 5.3 to 5.5 have been presented in the form of three manuscripts, looking further into the effect of age and risk of aggressive PCa. While Section 5.3 is a Review Article wherein we discussed the role of progressing age, and the various physical and gene-level changes that may affect the risk of PCa, Section 5.4 is a Research Article, wherein we have discussed our findings from the SNP genotype data from our cohort.

Beyond, we have also considered the SNP-SNP interactions and risk of aggressive PCa in our cohort in Section 5.5. Since the essence of genome mapping lies in the fact that the molecular markers need to be placed onto their respective positions on the genome for further analysis, we wanted to see how closely the SNPs are associated with one another with regards the risk of PCa in our cohort. Therefore, SNP-SNP interaction analysis was carried out by us. Although, fine-mapping of the genes could have been investigated, but owing to the time constraint, it was not feasible. Therefore, we consider this as shortcoming.

Since the results of the SNP-SNP interactions have also indicated a role of age and risk of aggressive PCa, this has also been categorized with our other findings pertaining to age-related risk factor for PCa.
Sections 5.3, 5.4 and 5.5 have been presented as published articles as follows.

**Prostate Cancer: Is It a Battle Lost to Age?**
Venkatesh Vaidyanathan, Nishi Karunasinghe, Anower Jabet, Radha Pallati, Chi Hsiu-Juei Kao, Alice Wang, Gareth Marlow and Lynnette R. Ferguson

**Effect of ageing and Single Nucleotide Polymorphisms associated with risk of aggressive prostate cancer in a New Zealand population**
Venkatesh Vaidyanathan, Vijay Naidu, Nishi Karunasinghe, Chi Hsiu-Juei Kao, Radha Pallati, Anower Jabet, Gareth Marlow, Prasanna Kallingappa, Lynnette R. Ferguson

**SNP-SNP interactions as risk factors for aggressive prostate cancer**
Venkatesh Vaidyanathan, Vijay Naidu, Nishi Karunasinghe, Anower Jabet, Radha Pallati, Gareth Marlow, Lynnette R. Ferguson
5.3. Prostate Cancer: Is It a Battle Lost to Age?

Abstract: Age is often considered an important non-modifiable risk factor for a number of diseases, including prostate cancer. Some prominent risk factors of prostate cancer include familial history, ethnicity and age. In this review, various genetic and physiological characteristics affected due to advancing age will be analyzed and correlated with their direct effect on prostate cancer.

5.3.1. Introduction

Aging is the irreversible process by which individuals undergo various physiological changes, and become vulnerable to various diseases, which in some cases can also be the cause of mortality (de Magalhaes, Curado et al. 2009; Birbrair, Zhang et al. 2013). Age is often considered an important non-modifiable risk factor for a number of diseases which include, but are not limited to neurodegenerative diseases such as age-specific Alzheimer’s disease (Jonsson, Atwal et al. 2012); cardiovascular diseases such as coronary heart disease, stroke (Jani and Rajkumar 2006; Finegold, Asaria et al. 2013), and hypertension (Pinto 2007); and certain cancers (Benz 2008; Brunet and Berger 2014) including prostate cancer (PCa) (Russo, Chen et al. 2012; Anderson, Sternberg et al. 2015).

PCa has one of the highest incidence rates amongst all diagnosed cancers in males worldwide (Cooper, McGuire et al. 2011). According to 2011 statistics, in males, PCa was the most common cancer registration in New Zealand, making up 27.3% of all male cancer registrations, and is a significant burden to the Public Health System (Karunasinghe, Lange et al. 2013). Some prominent risk factors of PCa include familial history, ethnicity and age (Bostwick, Burke et al. 2004). Only a very few cases of PCa are diagnosed in men less than 50 years old and around three-quarters of all the cases diagnosed are in men 65 years or older. For this review, we are considering 65 years as the cut-off (Nelen 2007). Certain other factors
which can also be related to PCa as risk factors are reproductive hormonal imbalances (Karunasinghe, Lange et al. 2013) and lifestyle factors such as, diet and smoking tobacco (Bostwick, Burke et al. 2004; Karunasinghe, Lange et al. 2013).

In this review, we will be looking into four parameters as risk for PCa which can be affected due to aging (Figure 5.5):

(i) Immunity and inflammatory response,
(ii) Cholesterol metabolism and obesity,
(iii) Free testosterone levels,
(iv) Genetic and epigenetic effects.

We will then correlate these changes with advancing age and try to establish if they have a direct effect on the onset and progression of PCa or not. By the end of this review we aim to answer three important questions with regards to the diagnosis and progression of PCa:

1. Is PCa a gerontological disorder fueled by an aging immune system, altered cellular metabolism, decreased levels of testosterone, and alterations in gene expression?
2. Is there potential for slowing or reverting these changes?
3. If the answer to question 1 is ‘yes’, can diagnostic tools distinguish aggressive PCa from non-aggressive PCa for possible early interventions?

To our knowledge, this is the first conclusive review to consider the alterations in risk factors for PCa with progressing age as a standard of variability in patients.
5.3.1.1. Risk Factors for Prostate Cancer

Most cancers are defined as complex diseases involving both environmental and genetic determinants as risk factors (Wu and Gu 2015). It is well documented that cancers of almost all the organs and organ systems can be caused due to environmental and lifestyle factors, including, and not restricted to, smoking tobacco, alcohol consumption, infectious agents, occupation, diet and obesity (Ferguson 2010). We suggest that the entire span of life spent pursuing the various lifestyle habits may have an accumulated effect on various factors such as immunity, inflammation, and even expression of the genes. Therefore, the role of aging is of much importance not only to understand the role it plays directly as a risk for diseases, but also indirectly.

With regards to PCa, it is well-established that there are three major risk factors. These are advancing age, ethnicity, and familial history (Gann 2002). Recent studies point out that alterations in genetic and epigenetic make-up are the basis for the development of malignancies (Orozco, Goh et al. 2013).

For most cancers, including PCa, however, the assessment of the exposures to external (or, environmental factors) and their effects still remains a challenge. Our understanding of the
The etiology of cancer in terms of environmental factors and genetic susceptibility is still rather limited, and the interplay among these etiological constituents is poorly understood (Wu and Gu 2015). For this review, as previously mentioned, we are focusing on the progressing age and its effect on the expression and progression of PCa.

5.3.1.2. Aging, Immunity, Inflammatory Response, and Prostate Cancer

Immune defense in higher vertebrates’ functions by the detection of a wide range of molecular patterns which are foreign to the tissues, inducing innate immunity and an inflammatory response (Medzhitov and Janeway 2000; Solana, Tarazona et al. 2012). The elderly exhibit an increase in susceptibility to various autoimmune, inflammatory, and/or infectious diseases (Hajishengallis 2010; Solana, Tarazona et al. 2012). Immunosenescence, or age-related alterations impairing the proper functioning of the immune system, is considered to be the major cause of most of the diseases associated with old age (Solana, Tarazona et al. 2012). Immunosenescence has an impact on both adaptive and innate immunity (Gomez, Nomellini et al. 2008; Hajishengallis 2014).

It is well-established that adaptive immunity declines with progressing age, due to factors such as reduced production of naïve T-cells, reduced diversity of antigen-recognition repertoire, alterations in signal transduction in T-cells with changes in the cytokine induction patterns, and reduction in expansion of clonal and functional specific T- and B-cells, as shown in Figure 5.6 (Weng 2006; Hajishengallis 2014). Aging of the thymus can also be linked to decreases in the production of naïve T-cells (Hajishengallis 2014). Alterations in innate immunity alone due to aging may not be the cause of immunodeficiency in individuals, but lead to dysregulation of immune response (Hajishengallis 2014). This dysregulation of immune response in aging individuals can eventually lead to either loss or gain of immune activity (Gomez, Nomellini et al. 2008; Hajishengallis 2010; Mahbub, Brubaker et al. 2011; Solana, Tarazona et al. 2012; Hajishengallis 2014). The gradual
decrease of naïve T-cells, accumulation of memory T-cells and the effector CD8⁺CD28⁻ T-cells, which are instrumental in over-secretion of pro-inflammatory cytokines, can lead to an imbalance in the pro- and anti-inflammatory networks (Franceschi, Bonafe et al. 2000; Meyer 2005; Weng 2006).

**Figure 5.6 Relation between aging and immunity.**

Inflammation can have a wide-spread effect on cancer, from development and progression of tumours, to response to therapies (Grivennikov, Greten et al. 2010). Chronic inflammatory responses induced by biological, chemical, physical, and/or mechanical injuries have been associated with a higher incidence rate of cancer in a wide range of human tissues (Thun, Henley et al. 2004; Grivennikov, Greten et al. 2010). Inflammation can be linked to cancer through two broad pathways, intrinsic and extrinsic (Mantovani; Sikora, Scapagnini et al. 2010). The intrinsic pathway of inflammatory response to cancer is initiated when oncogene activation is induced in the transformed cells, such that activation of transcription factor nuclear factor kappa-light-chain-enhancer of activated B-cells is affected by the production of inflammatory mediators (Sikora, Scapagnini et al. 2010). The extrinsic inflammatory response pathway caused due to infectious conditions also affects the risk of cancer.
development (Sikora, Scapagnini et al. 2010). Various risk factors, including environmental aspects are also identified to be associated with some kind of chronic inflammation (Grivennikov, Greten et al. 2010).

Some of the genes which play a common role as risk factors for PCa and inflammation pathways, include RNase L (RNASEL), Macrophage Scavenger Receptor 1 (MSR1), Glutathione S-Transferase P (GSTP1), Growth Differentiation Factor 15 (GDF15), Toll-like Receptor 4 (TLR4), Macrophage Inhibitory Cytokine-1 (MIC-1), Interleukin 1 Receptor Antagonist (IL1RN), Interleukin-8 (IL8), and Interleukin-10 (IL10) (De Nunzio, Kramer et al. 2011). Kazma, et al. (2015) investigated the association of 320 single nucleotide polymorphisms in 46 genes involved with the inflammation pathway with the risk of PCa in 494 patients with advanced diseases and 536 healthy men of African American and Caucasian ethnicities (Kazma, Mefford et al. 2012). Their results indicated that, although innate immunity and the inflammation pathway do not play a central role in the progression of PCa, they may play a role in the overall risk of the disease (Kazma, Mefford et al. 2012).

Anatomically, since the prostate gland is broadly classified into four lobes, some researchers have also looked into the expression and spread of cancer based on the individual lobe.

Das et al. (2013), while working on changes in proteomic profiles in the different lobes of male rats in different aging stages, have identified that aging in the dorsolateral and ventral lobes affected many proteins that are involved in vital cellular activities, especially inflammatory response (Das, Bortner et al. 2013). Of the many proteins, identified to be affected by aging, in each lobe, the expression of three proteins was identified to have increased, α-1 inhibitor 3, cysteine and glycine-rich protein, and ANXA1 (by the gene Annexin A1). Three proteins have also decreased expression uniformly in both lobes- hypoxia up-regulated protein 1, prolyl-1-hydroxyl-β peptide, and protein disulphide isomerase family
A, member 3 (Das, Bortner et al. 2013). These findings are interesting, as the role of genes is much more evident than the ageing of the organs as such.

We suggest that aging leads to a decline in immune response, and triggers the inflammatory pathways leading to development of PCa and these factors may not be playing a role in the progression of this disease but the risk of development of disease.

5.3.1.3. Aging, Cholesterol Metabolism, and Prostate Cancer

Effects of high fat diets (HFDs) on cholesterol metabolism have been associated with hyperlipidemia in humans (Shanmugasundaram, Visvanathan et al. 1986). HFDs increase the total and low-density lipoprotein (LDL) cholesterol levels in plasma, decrease high-density lipoprotein (HDL), and increase the total cholesterol to HDL ratio (Shanmugasundaram, Visvanathan et al. 1986). A life-long consumption of food rich in calories such as red meat, processed food substances (such as meat, dairy and fruit products) and/or high temperature cooking methods (Elmslie, Sellman et al. 2012) increases the chances of obesity and concentrations of triglycerides and cholesterol in serum. This may be due to increased energy intake compared to expenditure leading to a number of diseases such as cancer (Ferguson 2010; Elmslie, Sellman et al. 2012). High cholesterol in circulation has also been considered as a risk factor for solid malignancies, predominantly due to the upregulation of pathways such as cholesterol synthesis and/or inflammatory response (Zhuang, Kim et al. 2005).

Although cancers originating in different tissues may vary hugely in terms of overall type and etiology, they can still have common attributes of metabolic anomalies (Currie, Schulze et al. 2013). Cancer, at the cellular level, is a disease defined by uncontrolled cell growth and proliferation requiring cellular building blocks such as nucleic acids, proteins, and lipids (Currie, Schulze et al. 2013). Alteration in metabolism in cancer cells permits them to accumulate higher quantities of metabolic intermediates which can be used as building blocks in the body (Currie, Schulze et al. 2013). Detection of cholesterol deposits in tumour cells has
made it vital to analyse the cellular function of cholesterol and fatty acid metabolisms leading to the uncontrolled growth in these cells (Brown 2007; Currie, Schulze et al. 2013; Krycer and Brown 2013). Due to the uniqueness of the prostate gland, the association between cholesterol deposits and PCa has been identified to be very strong (Brown 2007; Prabhu, Krycer et al. 2013). Oxidation of fatty acids is also the source for an increase in the production of mitochondrial reactive oxygen species (ROS), which at high levels can be harmful to organelles, including the mitochondria (Rosca, Vazquez et al. 2012; Huang and Freter 2015), and various pathological states including cancers such as PCa (Dalleau, Baradat et al. 2013).

Various case-control (Platz, Leitzmann et al. 2006; Murtola, Tammela et al. 2007) and epidemiology studies (Shimizu, Ross et al. 1991) have demonstrated that individuals with elevated blood cholesterol levels run a higher risk of PCa (Krycer and Brown 2013). The epithelial cells of the prostate gland have two very unique features (Prabhu, Krycer et al. 2013). Prostate gland cells, owing to higher cholesterol synthesis as compared with the liver cells, have higher cholesterol levels than other tissues in the vicinity and this increase with aging and progression of PCa (Figure 5.7) (Brown 2007). Prostate epithelial cells also express higher level of glycolytic activity with reduced respiration (Ohvo-Rekila, Ramstedt et al. 2002; Prabhu, Krycer et al. 2013).
It is an inevitable fact that, when subjects undergo studies designed to identify the impact of amount and/or type of fat intake, there is an overall effect on the protein and/or carbohydrate in the diet as well (Lin, Aronson et al. 2015). Due to this, it is extremely difficult to pinpoint the effect that changes in diet have on the well-being of an individual (Lin, Aronson et al. 2015). It is therefore very important to look into details such as genes involved in metabolism to identify the regulation of fat breakdown and absorption. The effect of aging on these genes becomes crucial too, as aging not only means that the individuals have had a lifetime of certain dietary habits (which may cause higher oxidative stress and/or DNA damage), but also factors such as telomere shortening which leads to senescence (Eisenberg 2011).

5.3.1.4. Aging, Testosterone Level and Prostate Cancer

It is well established that the various symptoms of aging in males include and are not limited to fatigue, reduction in physical strength, lack of energy, lower and eventually loss of libido, reduced sexual performance, depression, and mood swings (Stanworth and Jones 2008). Various bodily compositions in healthy males also change with age-associated decline in free and bioavailable testosterone (Ellison, Bribiescas et al. 2002). Some examples of
alteration in body compositions include an increase in fat mass and decrease in muscle mass and bone mineral density (Denti, Pasolini et al. 1999; Vermeulen, Goemaere et al. 1999; Kenny, Prestwood et al. 2000; van den Beld, de Jong et al. 2000).

With testosterone being the primary androgen receptor-activating hormone identified in the Wolffian duct, and also responsible for the development of the primary sexual characteristics (Roy, Lavrovsky et al. 1999), it is the main hormone of interest in this review. The overall level of the biochemical testosterone also decreases with aging (Stanworth and Jones 2008). What is interesting from the point of research is that the symptomatic evidence suggests normal aging in males to be very similar to those with mild androgen deficiency (Stanworth and Jones 2008). In concordance with the core idea of this review, aging in males and the effect on testosterone levels in males with PCa is another important aspect.

There is a substantial amount of epidemiological data to prove that serum free and total testosterone, and adrenal steroid dehydroepiandrosterone (DHEA) levels decrease with the normal aging process after peaking reaching the age bracket of 20–30 years (Harman, Metter et al. 2001; Feldman, Longcope et al. 2002; Arnold 2009). Examples of the decrease in the levels of serum testosterone with aging have been cited in cross-sectional and longitudinal studies (Harman, Metter et al. 2001; Wu, Tajar et al. 2008). On the contrary, certain other androgen axis products such as luteinizing hormone, follicular stimulating hormone and dihydrotestosterone levels, and sex hormone-binding globulin levels have been identified to increase with age (Feldman, Longcope et al. 2002; Araujo and Wittert 2011). Reduction of testosterone could also be at least partly due to it’s being aromatized to estrogen with increasing age (Araujo and Wittert 2011). According to Bélanger, et al. (1994), a decline in adrenal DHEA is responsible for up to a 50% reduction in total androgens in men beyond the age of 40 years (Belanger, Candas et al. 1994).
Various circulating androgens, including testosterone, have important roles in the growth of the prostate gland, and PCa (Shin, Hwang et al. 2010). The specific pathway mechanism or the role of testosterone level on progression of PCa, however, is still to be well defined (Shin, Hwang et al. 2010). The level of circulating testosterone in men more than 45 years of age, across various ethnicities, shows a tendency to converge down as compared with vastly varied levels in men of age less than 30 years (Ellison, Bribiescas et al. 2002). Many researchers have shown that higher levels of circulating testosterone do not correlate with an increased risk of PCa (Hoffman, DeWolf et al. 2000; Stattin, Lumme et al. 2004). However, low levels of serum testosterone do have a direct correlation with progression of the disease to aggressive PCa (Kumar, Wadhwa et al. 1990; Hoffman, DeWolf et al. 2000; Schatzl, Madersbacher et al. 2001; Shin, Hwang et al. 2010). Moreover, with a life-long consumption of fatty food substances, the levels of testosterone can also be affected by this. These absolutely contradicting effects on the levels of testosterone can also cause hormonal imbalance, as shown in Figure 5.8.
These observations suggest that variation in the pattern of declining levels of free testosterone with progressing age is primarily caused by reproductive physiological variations in various populations and not because of the ethnic differences in populations (Ellison, Bribiescas et al. 2002). It is also important to mention that the drastic depletion of testosterone levels in Caucasian populations from elevated levels in young-adulthood to comparatively much lower levels in adult-aged men causes drastic alterations in hormonal balance which may be a potential cause for higher risk of PCa (Ellison, Bribiescas et al. 2002).

Munetomo et al. (2015) have shown an increase in androgen receptor (AR) expression in the hypothalamus of rats as they age (3 v/s 24 months) (Munetomo, Hojo et al. 2015). Meanwhile, Pomerantz et al. (2015) have also shown extensive reprogramming of AR cistrome during prostate epithelial transformation into tumours (Pomerantz, Li et al. 2015).
The close correlation and co-activation between AR and telomeres have also been discussed in detail by Zhou et al. (2013) (Zhou, Richardson et al. 2013). Therefore, aging related telomere shortening (Hurwitz, Heaphy et al. 2014) could also have direct effects on normal functioning of the AR.

5.3.1.5. Aging, Genetic and Epigenetic Effects and Prostate Cancer

One of the other aspects linking alteration of testosterone levels, aging and PCa, and thereby, worthy of mention in this review, is alterations of DNA methylation. Ammerpohl, et al. (2013), suggested that androgens affect sexual dimorphism in humans and thereby change DNA methylation marks in the epigenome (Ammerpohl, Bens et al. 2013). Interestingly, on one hand hypermethylation of certain genes such as GSTP1 is well established to be consistent with the transition of PCa from intraepithelial neoplasia stage to becoming a frank carcinoma (Kwabi-Addo, Chung et al. 2007), and on the other, certain other studies exhibit that DNA methylation and histone modifications generally recapitulate the normal aging process (Osorio, Varela et al. 2010). Therefore, it is very difficult to specifically identify the role of aging on methylation changes, leading to and/or progression of PCa.

Drastic changes are observed in epigenetic patterns during growth and development; most of these events are biologically programmed and absolutely necessary for healthy being. However, changes in the epigenome in mature (adult) somatic tissues mirror aging-associated deleterious effects (Jung and Pfeifer 2015). Alterations in biological processes, cellular responses, and disease states, are all parameters well-established to have an association with changes in gene expression (de Magalhaes, Curado et al. 2009). Many microarray studies have been carried out to define the process of aging and to identify potential genes, gene expressions, and biomarkers of risk factors for many gerontological diseases (Ida, Boylan et al. 2003; de Magalhaes, Curado et al. 2009). However, aging gene expression studies have a number of complications. Two of the main issues faced by research in this area are the fact
that only a few genes are identified as being differentially expressed, and also, fewer genes are found to overlap with the effect in a wide range of tissues (Weindruch, Kayo et al. 2002; de Magalhaes, Curado et al. 2009). Rodwell, et al. (2004), suggested that a very small proportion of transcriptional response is tissue-specific, and therefore molecular signatures of aging may overall be identified even in unrelated tissues (Rodwell, Sonu et al. 2004). These signatures, however, can be subject to varied interpretations rather than an active aging program (Rodwell, Sonu et al. 2004; de Magalhaes, Curado et al. 2009). de Magalhes, et al. (2009), were able to integrate gene expression analyses from various studies to identify genes that have a tendency to over- and/or under-express with progressing age (de Magalhaes, Curado et al. 2009). The inflammatory response pathway is one of the most important pathways that is upregulated with aging in humans (de Magalhaes, Curado et al. 2009). Over-expression of anti-apoptotic genes and cell-cycle regulators such as Granulin (GRN), annexins, and genes playing a protective role during oxidative stress and detoxification of lipid peroxidation of end products such as Glutathione S-Transferase-1 (GST1) are commonly observed to be directly related to progressing age (de Magalhaes, Curado et al. 2009). Genes that under-express with aging are not only fewer than the over-expressing genes, but are also simpler to interpret, as they are predominantly identified in the energy metabolism categories such as cholesterol metabolism (de Magalhaes, Curado et al. 2009). Negative regulation of transcription strongly suggests that transcriptional activities decrease with aging, thus supporting the hypothesis that RNA synthesis decreases with aging (de Magalhaes, Curado et al. 2009). Nevertheless, the total protein content of an individual may not alter with age always; therefore, decreasing mRNA may lead to the accumulation of proteins with anomalies (de Magalhaes, Curado et al. 2009).

The bulk of the mammalian genome gets transcribed to non-coding ribonucleic acids (RNAs) (Thum 2014; Devaux, Zangrando et al. 2015). Two major groups of non-coding
RNAs that play important roles as epigenetic regulators of gene expression are long non-coding RNAs (lncRNAs) (Devaux, Zangrando et al. 2015) which are comprised of nucleotide sequences >200 bp and small non-coding RNAs which include microRNAs (miRNAs) (Jung and Suh 2012), comprised of nucleotide sequences <200 bp. Certain molecular mechanisms permit lncRNAs to regulate cellular and tissue functions either by activating or suppressing gene expression (Devaux, Zangrando et al. 2015), whereas miRNAs usually down-regulate gene expression either by mRNA degradation or by blocking off translation (Guo, Ingolia et al. 2010; Jung and Suh 2012). The mechanisms of action of long non-coding RNA with regards to fat metabolism, diabetes, diseases of mitochondrial dysfunction, age-associated muscle pathology, cancer, declining stress response, age-associated immune decline and age-associated neurodegeneration should also be looked into (Lopez-Otin, Blasco et al. 2013). A link between miRNAs and longevity has already been demonstrated in *Caenorhabditis elegans*, thereby implicating the vital role of miRNAs in the regulation of lifespan and the overall process of aging (Jung and Suh 2012). miRNAs have a higher impact as gene modifiers compared to lncRNAs because not only can a single miRNA target multiple mRNAs, but one mRNA can also be targeted by multiple, very different miRNAs (Jung and Suh 2012).

Hooten, *et al.* (2010), used mononuclear cells obtained from peripheral blood to estimate the expression level of miRNA in young and old populations, and identified a negative correlation with advancing age (Noren Hooten, Abdelmohsen et al. 2010). This information is very important to define the possibility of miRNA playing a crucial role in the process of aging. miRNAs have also been identified to be key players in controlling metabolic homeostasis and related diseases in individuals (Rottiers and Naar 2012).

Owing to the unique expression patterns in tissues affected with cancer, miRNA profiling in PCa is a common approach adapted to diagnostic, progressive and/or therapeutic use.
Certain miRNAs have been identified to be associated with PCa. One such example is miR-21 (Amankwah, Anegbe et al. 2013). miR-21 has also been identified to have increased expression in the liver biopsies of obese individuals (Vinciguerra, Sgroi et al. 2009). This correlation between the role and abundant expression of miR-21 is important to understand the overall relation between obesity (which can be age-related) and PCa. Target and pathway analysis for serum age-associated miRNAs also explain the role played by miRNAs on PCa due to aging of an individual. We recently proposed that certain miRNAs can be begotten from food sources that may contribute to obesity and also to PCa (Vaidyanathan, Krishnamoorthy et al. 2016). Hooten, et al. (2013) also used TargetScan 6.2 to predict that miR-151a-3p, miR-181a-5p and miR-1248, targeted 115, 626 and 265 mRNAs respectively (Noren Hooten, Fitzpatrick et al. 2013). A number of neurological diseases and cancer, including PCa were identified to overlap among these three miRNAs thereby proving that age is a significant risk factor for certain cancers including PCa (Noren Hooten, Fitzpatrick et al. 2013).

5.3.2. Discussion

The present review article targets two very important questions in the diagnosis of PCa- is this disease caused due to age-dependent immunosenescence, aging cholesterol metabolism, and androgen axis, or does failing genetics play the most crucial role in the expression and progression of PCa. The four parameters that have been followed in this review, namely changes in immunity and inflammatory response, cholesterol metabolism and obesity, effects on free testosterone level, and variations in gene expression, are all inter-linked to each other with regards to their tendency to impact expression and progression of PCa in humans with advancing age. These four parameters have isolated and combined effects on the expression and progression of PCa.
Advancing age of patients, for example, alters the ability of certain genes to express and push towards an overall tendency to over- or under-express. The expression levels of certain genes, such as Glutathione Transferases, have previously been identified to have a direct relation to progressing age, and these genes in turn affect the expression of diseases such as PCa through various pathways such as inflammatory response. Interestingly, a similar chain of events pertaining to PCa can be observed when considering the effect on fat-mass and obesity-associated gene (Lewis, Murad et al. 2010) initiated by progressing age. The effect of progressing age on switching on and off genes, especially on post-transcriptional gene modifiers such as lncRNAs and miRNAs are also well documented (Kanherkar, Bhatia-Dey et al.). This is a major change as well, bearing in mind that miRNAs, constituting only around 1% of the entire genome have been estimated to target as many as 30% of the genes (Lewis, Burge et al. 2005). These direct our understanding of PCa towards the conclusion that it is in fact a gerontological disorder that gets aggravated with alterations in the expression of genes that govern the immune system, cellular metabolism and testosterone production. Stromal weakness with age adds to this problem, providing a platform for easy invasion of the malignant cells.

Since these changes occur at the molecular level, it will be very interesting to see if they can be reversed using modern biological tools such as siRNA technology or clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR associated protein 9 (CRISPR/Cas9) targeted genome editing, accompanied with lifestyle changes. Therefore, there is a need to be able to ablate the genes in cells so that we can understand what the loss of function would look like and to have a model where we can add back the protein in the absence of confounding local production. This situation has changed recently, however, and it is now possible to achieve very high efficiency gene targeting using the CRISPR/Cas9 technology (Mali, Yang et al. 2013; Ran, Hsu et al. 2013). As a result, it is now possible to
generate somatic gene knockout to produce isogenic cancer cell models that possess or lack the genes with abnormal expression and use the synthetic protein to study its mechanism of action. Even if highly targeted CRISPR/Cas work in humans requires extensive work into the future, this technology can be used in in vitro models to understand downstream pathways modified by such gene deletions. Such knowledge can be used in future targeted therapies. Until then, however, molecular techniques such as genotyping should be used in parallel with the pathophysiological details.

This brings us to the third question being addressed in this review—can diagnostic tools identify aggressive PCa from non-aggressive PCa. It is recognized that aging cannot be reversed, but, genetic tests such as association of various clinical characteristics with single nucleotide polymorphisms (SNPs) can be used for early diagnosis and personalized treatment (Karunasinghe, Lange et al. 2013; Yu, Huang et al. 2013; Karunasinghe, Zhu et al. 2016). The use of prostate-specific antigen (PSA) levels in PCa diagnosis is still controversial (Barry 2009). A strong statement by the United States Preventive Services Taskforce in 2008 and 2012 against PSA testing, as well as those of other national bodies (2008; Moyer 2012; Bell, Connor Gorber et al. 2014), has thrown the use of PSA testing into doubt, resulting in confusion amongst patients and their health practitioners. Recent studies indicate the consequences following the withdrawal of PSA screening which has resulted in an overall increase in metastatic PCa incidence (Weiner, Matulewicz et al. 2016). According to these authors, the relative increase in metastatic PCa incidence compared to data of 2004 was highest in the age range of 55–69 years with an increase by 92%.

Various genetic studies with aggressive PCa have been carried out (Duggan, Zheng et al. 2007; Amin Al Olama, Kote-Jarai et al. 2013). Such studies can pave way for early differentiation of men that are more likely to develop aggressive disease. Such information will support precautionary lifestyle changes for at risk men as well as differentiation of those
that required early interventions for aggressive PCa treatments. Various genome-wide association studies (GWAS) have also been carried out to identify the possibility of aggressiveness of PCa associated with SNPs (Easton and Eeles 2008; Knipe, Evans et al. 2014), but a well-documented database is needed to account for ethnic variations. GWAS provide the tools to identify common and low-penetrance loci of diseases, such as PCa, without prior knowledge of the location and/or function (Easton and Eeles 2008). The database thereby created, will aid in identify the important SNPs and the effect of external factors, including and not limited to aging, downstream of the genes harboring the SNPs. Data sharing at this level will also be of much help for researchers to understand gene x gene, gene x environment and gene x diet interactions. Race is another risk factor especially African-American, and in almost all research carried out, it is corrected for (Kinseth, Jia et al.), but the use of database thus created can also help researchers better understand aging process, if different between African American and Caucasian with PCa, which is not consistent even in GWAS (Bensen, Xu et al. 2013), and thus help diagnose and control the progression of the disease. It will also be interesting to see if an approach targeting SNPs in genes specific to immunity and inflammatory response, cholesterol metabolism and obesity, and androgen metabolism is carried out to check for association with progression of PCa in patients with and without the aggressive form of the disease.
5.4. Effect of ageing and Single Nucleotide Polymorphisms associated with risk of aggressive prostate cancer in a New Zealand population

Abstract: Prostate cancer is one of the most significant male health concerns worldwide, various researchers carrying out molecular diagnostics have indicated that genetic interactions with biological and behavioral factors play an important role in the overall risk and prognosis of this disease. Single nucleotide polymorphisms are increasingly becoming strong biomarker candidates to identify susceptibility of prostate cancer. We carried out risk association of different stages of prostate cancer to a number of single nucleotide polymorphisms to identify the susceptible alleles in a New Zealand population and checked the interaction with environmental factors as well. We have identified a number of single nucleotide polymorphisms to have associations specifically to the risk of prostate cancer and aggressiveness of the disease, and also certain single nucleotide polymorphisms to be vulnerable to the reported behavioral factors such as tobacco smoking, consumption of alcohol, among others.

5.4.1. Introduction

Prostate cancer (PCa) is one of the most significant non-skin cancer male health concerns worldwide (Vaidyanathan, Naidu et al. 2017). Moreover, it is estimated that at least 1 in 6 PCa patients is at risk of developing aggressive PCa (Cooperberg, Vickers et al. 2010). These are very alarming statistics. The identification of a predictive biomarker and/ or treatment of this disease is therefore of much importance, more so from the New Zealand point of view, because the highest rate of recording of men with PCa, relative to the population of men, is observed in the Oceania region (Jemal, Bray et al. 2011; Karunasinghe, Han et al. 2012). With various biological and behavioral factors established as playing crucial role in the

Although, age, ethnicity, and family history are the three most widely accepted risk factors for PCa (Gann 2002; Bostwick, Burke et al. 2004; Karunasinghe, Lange et al. 2013), yet nothing much can clinically be done to alter or reverse the effect of these on human health and immunity. Of these three risk factors, age is the most significant risk factor for aggressive PCa (Haas and Sakr 1997; Vaidyanathan, Karunasinghe et al. 2016). In the same line, we believe that healthy ageing, can control the expression of the aggressive form of this disease.

We recently identified gene x environment interaction(s) and the risk of aggressive PCa in a New Zealand population and defined a trend that certain lifestyle habits and effects such as tobacco smoking, and high body mass index (BMI), also have an influence on the aggressiveness of the disease (Vaidyanathan, Naidu et al. 2017). Even with progressing age, which cannot be curtailed, certain lifestyle habits may stay put. Here, we employed some statistical tools and analyzed data generated by genotyping single nucleotide polymorphisms (SNPs) of interest to understand the effect of ageing on external factors and effects such as tobacco smoking, alcohol consumption; and high BMI and risk of aggressive PCa.

Here we present the analysis of the data obtained following the genotyping of 138 SNPs, using SEQUENOM MassArray iPLEX® assay and TaqMan® SNP genotyping procedures in a New Zealand cohort. The cohort includes New Zealand men of self-declared European ethnicity with different clinically diagnosed grades/stages of PCa, and gender matched healthy controls within similar age range. We have identified the association of SNPs as risk for aggressive PCa as well as the influence of external factors including age in risk
modification. This, we believe, is the first such study on genetic and environmental risk association with ageing and risk of aggressive PCa in a New Zealand cohort.

5.4.2. Materials and Methods

5.4.2.1. Study population

The study cohort is the same as in section 5.1.2.1. Because of the influence of age in this disease (Vaidyanathan, Karunasinghe et al. 2016), care was taken to invite men between the age categories of 40 to 90 years (at the time of diagnosis for patients with PCa and at the time of recruitment for healthy controls) to participate in this study. We have considered men more than 65 years of age as elderly or older person, as per the norms of World Health Organization (WHO).

5.4.2.2. Definition of aggressiveness:

The definition of the aggressiveness of PCa has already been mentioned in Section 5.1.3.1.

5.4.2.3. Statistical analysis:

The statistical analysis was similar to the one explained in the Section 5.1.3.2 of this thesis.

5.4.3. Results

5.4.3.1. Age, Pathology, BMI and lifestyle:

The median age of our aggressive PCa cohort was 65.6 y (with the 75th percentile being 71.0 y), the median age of our non-aggressive PCa cohort was 67.3 y (with the 75th percentile being 71.8 y) and the median age of our healthy controls cohort was 57.0 y (with the 75th percentile being 66.0 y). We identified that the age is statistically significant between
aggressive PCa and healthy controls, and between non-aggressive PCa and healthy controls, but not between the aggressive PCa and non-aggressive PCa cases.

Since the main aim of this article is to identify the role of ageing and statistically adjusting for this parameter in isolation and in combination with various demographic factors such as alcohol consumption, smoking tobacco, and with levels of obesity among the patients recruited for our study have been mentioned in Section 5.1.4.2. We are presenting the data for variation in age as risk for PCa in Table 5.9, Table 5.10, and Table 5.11.

### Table 5.9 Association between age and aggressive prostate cancer vs healthy controls.

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Pathology</th>
<th>N'</th>
<th>Percentage of men ≥65 years</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive vs Healthy Control</td>
<td>Aggressive</td>
<td>90</td>
<td>107</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Healthy Control</td>
<td>266</td>
<td>103</td>
<td>369</td>
<td></td>
</tr>
<tr>
<td>Agressive vs Healthy Control Aggressive</td>
<td></td>
<td>54.31%</td>
<td></td>
<td>3.070334</td>
<td>7.98E-10</td>
</tr>
<tr>
<td></td>
<td>Healthy Control</td>
<td>27.91%</td>
<td></td>
<td>(2.1399 – 4.4052)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Median age of our aggressive PCa cohort was 65.6 (75th percentile- 71.0)

Median age of our healthy controls cohort was 57.0 (75th percentile being 66.0)

**Aggressive PCa cohorts vs healthy controls: p-value <0.001**

### Table 5.10 Association between age and aggressive prostate cancer vs non-aggressive prostate cancer.

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Pathology</th>
<th>N'</th>
<th>Percentage of men ≥65 years</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggressive vs Non-Aggressive</td>
<td>Aggressive</td>
<td>90</td>
<td>107</td>
<td>197</td>
<td>0.642643</td>
</tr>
<tr>
<td></td>
<td>Non-Aggressive</td>
<td>20</td>
<td>37</td>
<td>57</td>
<td>0.173763</td>
</tr>
</tbody>
</table>
| Median age of our non-aggressive PCa cohort was 67.3 (75th percentile- 71.8)

Median age of our aggressive PCa cohort was 65.6 (75th percentile- 71.0)

**Non-aggressive PCa cohorts vs aggressive PCa cohort: p-value =1.0**
Table 5.11 Association between age and non-aggressive prostate cancer vs healthy controls.

<table>
<thead>
<tr>
<th>Compared groups</th>
<th>Pathology</th>
<th>N’</th>
<th>Percentage of men ≥65 years</th>
<th>OR (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Aggressive vs Healthy Control</td>
<td>Non-Aggressive</td>
<td>20</td>
<td>37</td>
<td>57</td>
<td>64.91%</td>
</tr>
<tr>
<td></td>
<td>Healthy Control</td>
<td>266</td>
<td>103</td>
<td>369</td>
<td>27.91%</td>
</tr>
</tbody>
</table>

Median age of our non-aggressive PCa cohort was 67.3 (75th percentile- 71.8)

Median age of our healthy controls cohort was 57.0 (75th percentile being 66.0)

Non-aggressive PCa cohorts vs healthy controls: p-value <0.001

Tables 5.9- 5.11 legend: N’= number; OR= Odds Ratio; 95% CI= 95% confidence interval

5.4.3.2. Genetic polymorphism variations and risk of prostate cancer:

The tables show the results of the statistically significant SNPs associated with risk of PCa between patients with aggressive PCa and healthy controls (Table 5.12), between patients with aggressive and non-aggressive PCa (Table 5.13), and patients with non-aggressive PCa and healthy controls (Table 5.14), all assessed before and after the adjustment for various demographic parameters with and without age aspect. Variations in the tested allele between patients recruited for this study with aggressive PCa, non-aggressive PCa and healthy controls for all the SNPs irrespective of statistical significance have been included in Supplementary Tables 1a and 1b and 2 and are available online.
Table 5.12 Statistically significant SNP associated with gene x environment effect on risk of aggressive prostate cancer v/s healthy controls after adjusting for each environmental parameter individually and along with age

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Gene location</th>
<th>SNP ID</th>
<th>Tested allele</th>
<th>Gene name</th>
<th>Before any adjustment</th>
<th>After adjustment for BMI</th>
<th>Tobacco smoking</th>
<th>Alcohol consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1p22.3</td>
<td>rs845</td>
<td>C</td>
<td>SEP15</td>
<td>p = 0.04139</td>
<td>OR = 1.6217</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2p23.1</td>
<td>rs632148</td>
<td>C</td>
<td>SRD5A2</td>
<td>p = 0.02485</td>
<td>OR = 1.642</td>
<td>p = 0.03795</td>
<td>OR = 1.572</td>
</tr>
<tr>
<td>3</td>
<td>3p22</td>
<td>rs1799977</td>
<td>A</td>
<td>MLH1</td>
<td>p = 0.01607</td>
<td>OR = 1.547</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3p22</td>
<td>rs1799977</td>
<td>A</td>
<td>PPARG</td>
<td>p = 0.000173</td>
<td>OR = 4.534</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6p21.3</td>
<td>rs1024432</td>
<td>A</td>
<td>CCHCR1</td>
<td>p = 0.02344</td>
<td>OR = 1.383</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7q33</td>
<td>rs130667</td>
<td>T</td>
<td>LEP</td>
<td>p = 0.02866</td>
<td>OR = 1.572</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17q21</td>
<td>17q25</td>
<td>11q12</td>
<td>11q13</td>
<td>10p15</td>
<td>10q11</td>
<td>8q24</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>rs887391</td>
<td>rs799923</td>
<td>rs6502051</td>
<td>rs2722720</td>
<td>rs11228565</td>
<td>rs10896438</td>
<td>rs7931342</td>
<td>rs12529</td>
<td>rs7920517</td>
</tr>
<tr>
<td>SLC26A6</td>
<td>BRCA1</td>
<td>FASN</td>
<td>FADS2</td>
<td>MYEOV</td>
<td>AKR1C3</td>
<td>MSMB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.005094</td>
<td>0.04184</td>
<td>0.02189</td>
<td>0.002322</td>
<td>0.0007423</td>
<td>0.04685</td>
<td>0.01227</td>
<td>0.02883</td>
<td></td>
</tr>
<tr>
<td>1.594</td>
<td>1.525</td>
<td>1.433</td>
<td>1.4985</td>
<td>1.565</td>
<td>1.294</td>
<td>1.400</td>
<td>1.883</td>
<td></td>
</tr>
<tr>
<td>0.01177</td>
<td>0.04996</td>
<td>0.009652</td>
<td>0.0007877</td>
<td>0.001017</td>
<td>0.006002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.845</td>
<td>1.329</td>
<td>1.9872</td>
<td>1.6231</td>
<td>1.6</td>
<td>1.4729</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01913</td>
<td>0.01887</td>
<td>0.001213</td>
<td>0.000222</td>
<td>0.005373</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.744</td>
<td>1.8382</td>
<td>1.5810</td>
<td>1.682</td>
<td>1.4679</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.009695</td>
<td>0.04415</td>
<td>0.01579</td>
<td>0.000432</td>
<td>0.000271</td>
<td>0.002065</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.892</td>
<td>1.346</td>
<td>1.90439</td>
<td>1.6806</td>
<td>2.2742</td>
<td>1.5615</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02494</td>
<td>0.04382</td>
<td>0.01279</td>
<td>0.001449</td>
<td>0.00637</td>
<td>0.0378</td>
<td>0.01025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.702</td>
<td>1.5384</td>
<td>1.9080</td>
<td>1.5669</td>
<td>1.612</td>
<td>1.3049</td>
<td>1.4196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01295</td>
<td>0.01067</td>
<td>0.0006136</td>
<td>0.000783</td>
<td>0.005115</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.84</td>
<td>1.9704</td>
<td>1.6463</td>
<td>1.626</td>
<td>1.4861</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.04069</td>
<td>0.04425</td>
<td>0.001049</td>
<td>0.001481</td>
<td>0.0258</td>
<td>0.004713</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.653</td>
<td>1.7418</td>
<td>1.61134</td>
<td>1.593</td>
<td>1.3424</td>
<td>1.5035</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02386</td>
<td>0.03186</td>
<td>0.03105</td>
<td>0.000411</td>
<td>0.001756</td>
<td>0.00248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.775</td>
<td>1.392</td>
<td>1.8549</td>
<td>1.7070</td>
<td>1.598</td>
<td>1.5710</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome</td>
<td>Position</td>
<td>Gene</td>
<td>SNP ID</td>
<td>Minor Allele</td>
<td>Minor Allele Frequency</td>
<td>Standardized Beta</td>
<td>P-value</td>
<td>OR 95% CI</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>--------</td>
<td>--------------</td>
<td>------------------------</td>
<td>-------------------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>17</td>
<td>21162652</td>
<td>KLK3</td>
<td>rs2659122</td>
<td>C</td>
<td>0.01748</td>
<td>1.5</td>
<td>0.04748</td>
<td>1.345</td>
</tr>
<tr>
<td>18</td>
<td>17680852</td>
<td>MMP9</td>
<td>rs17632542</td>
<td>G</td>
<td>0.008268</td>
<td>1.1</td>
<td>0.04928</td>
<td>1.5163</td>
</tr>
<tr>
<td>19</td>
<td>20518531</td>
<td>NUPT1</td>
<td>rs3918256</td>
<td>G</td>
<td>0.00749</td>
<td>1.4</td>
<td>0.04894</td>
<td>1.3116</td>
</tr>
<tr>
<td>20</td>
<td>Xp11</td>
<td>NUPT1</td>
<td>rs5945619</td>
<td>T</td>
<td>0.005749</td>
<td>1.4</td>
<td>0.005749</td>
<td>1.694</td>
</tr>
<tr>
<td>Sl. No.</td>
<td>Gene location</td>
<td>SNP ID</td>
<td>Tested allele</td>
<td>Gene name</td>
<td>Before any adjustment</td>
<td>After adjustment for Age</td>
<td>After adjustment for BMI</td>
<td>After adjustment for BMI + Age</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>--------</td>
<td>--------------</td>
<td>-----------</td>
<td>----------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>1</td>
<td>19q13</td>
<td>rs887391</td>
<td>C</td>
<td>SRD5A2</td>
<td>0.01731</td>
<td>1.799</td>
<td>0.0124</td>
<td>2.121</td>
</tr>
<tr>
<td>2</td>
<td>9q33.1</td>
<td>rs11536889</td>
<td>A</td>
<td>MLPH</td>
<td>0.0263</td>
<td>2.303</td>
<td>0.02702</td>
<td>1.799</td>
</tr>
<tr>
<td>3</td>
<td>7q33</td>
<td>rs10244329</td>
<td>T</td>
<td>PODXL</td>
<td>0.03126</td>
<td>2.062</td>
<td>0.03816</td>
<td>1.801</td>
</tr>
<tr>
<td>4</td>
<td>7q32</td>
<td>rs373035</td>
<td>T</td>
<td>LEP</td>
<td>0.03222</td>
<td>1.621</td>
<td>0.03593</td>
<td>1.674</td>
</tr>
<tr>
<td>5</td>
<td>2q37.2</td>
<td>rs2292884</td>
<td>G</td>
<td>TLR4</td>
<td>0.02614</td>
<td>1.801</td>
<td>0.03623</td>
<td>2.086</td>
</tr>
<tr>
<td>6</td>
<td>2p23.1</td>
<td>rs632148</td>
<td>C</td>
<td>LEP</td>
<td>0.0251</td>
<td>1.621</td>
<td>0.03528</td>
<td>1.674</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>19q13.33</td>
<td>rs7633242</td>
<td>T</td>
<td>KLK3</td>
<td>0.04647</td>
<td>3.194</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>20q13.12</td>
<td>rs3918256</td>
<td>A</td>
<td>MMP9</td>
<td>0.04959</td>
<td>1.555</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.9200</td>
<td>1.0324</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8010</td>
<td>1.0220</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8925</td>
<td>1.0267</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8989</td>
<td>1.0269</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.9456</td>
<td>1.0260</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8731</td>
<td>1.0269</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.8731</td>
<td>1.0269</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.9215</td>
<td>1.0269</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.14 Statistically significant SNP associated with gene x environment effect on risk of non-aggressive prostate cancer v/s healthy controls after adjusting for each environmental parameter individually and along with age.

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Gene location</th>
<th>SNP ID</th>
<th>Tested allele</th>
<th>Gene name</th>
<th>Before any adjustment</th>
<th>After adjustment for Age</th>
<th>BMI</th>
<th>After adjustment for BMI</th>
<th>After adjustment for BMI + Age</th>
<th>Tobacco smoking</th>
<th>After adjustment for Tobacco smoking</th>
<th>Alcohol consumption</th>
<th>After adjustment for Alcohol consumption + Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2q37.2</td>
<td>rs2292884</td>
<td>C</td>
<td>MLPH</td>
<td>0.02375</td>
<td>1.774</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7q32</td>
<td>rs3735035</td>
<td>C</td>
<td>PODXL</td>
<td>0.03493</td>
<td>1.572</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9q33.1</td>
<td>rs11536889</td>
<td>C</td>
<td>TLR4</td>
<td>0.02727</td>
<td>1.801</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>15q26.3</td>
<td>rs465373</td>
<td>A</td>
<td>SEPS1</td>
<td>0.04213</td>
<td>1.801</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tables 2.1-2.3 colour legends risk association:

SNPs statistically significantly associated with risk of aggressive PCa across various classifications both, before and after adjusting for the environmental and age parameters
5.4.4. Discussion

It is well-established that there are three major risk factors for PCa, namely, advancing age, ethnicity, and familial history (Gann 2002). Recent studies indicate alterations in genetic and epigenetic make-up as the basis for the development of various malignancies (Orozco, Goh et al. 2013) and is in line with our findings with regards risk of aggressive PCa (Vaidyanathan, Naidu et al. 2017). In the current article, the data obtained by SNP genotyping and reported in Vaidyanathan et al., (2017) (Vaidyanathan, Naidu et al. 2017) was further analyzed to identify risk association with aggressive PCa with the effect of non-genetic or environmental factors after being adjusted statistically with and without the influence of ageing on them.

Out of the 97 SNPs studied by us, only 5 SNPs were identified to be significantly associated with risk of aggressive PCa when compared with healthy control across all combinations before and after adjustment, 4 SNPs were significantly associated with risk of aggressive PCa when compared with non-aggressive PCa across all combinations before and after adjustment, and no SNPs were identified to be significantly associated with risk of non-aggressive PCa compared to healthy controls across all combinations before and after adjustment.

Although the genome-wide association studies (GWAS) are used for the identification of the direct role SNP association plays as for aggressive PCa, yet we believe that SNP interactions with demographic and lifestyle factors could also add to the allelic effect producing a modified risk of a disease. These SNPs identified herewith to have come up significant could be indicating a unique situation for New Zealand men with PCa, and can be used as a model for other chronic diseases.
5.4.4.1. Age at diagnosis and age at recruitment (prostate cancer patients and healthy controls respectively) and risk of prostate cancer:

Age is a major risk factor for PCa, as reported (Karunasinghe, Han et al. 2013; Vaidyanathan, Karunasinghe et al. 2016). However, in the data presented in our present study we did not consider the role of ageing, as we wanted to see the effect of gene and environment aspects in the expression and progression of PCa. Age, being irreversible, but other environmental factors being more under one’s control we focused on those aspects to identify any link and define the means by which high-risk PCa can be controlled.

We found statistical correlation of age as risk for aggressive PCa when compared to healthy controls. It is often suggested that older men (≥65 years of age), especially if they have not undergone PSA testing, are more likely to develop the aggressive form of PCa, if they develop PCa, and are also more likely to die of the same as compared to younger men (≤64 years of age) (Vellekoop and Loeb 2013; Tosoian, Alam et al. 2017). This indication of age as a risk factor for aggressive PCa is in line with the findings in our cohort as well (Table 5.9). Consistent with the findings of other groups, we found that age of an individual is associated with risk of non-aggressive PCa too when compared with healthy controls (Table 5.11), but has no significant correlation as risk for aggressive PCa when compared to patients with non-aggressive PCa (Table 5.10), as is understandable. Diseases such as PCa often have an onset with progressing age (Kelly, Rosenberg et al. 2017), but the aggressiveness may not be solely age-dependent (Vaidyanathan, Naidu et al. 2017).

It is also worthy of mention that it is often stated that young men (≤55 years of age) who are diagnosed for PCa, are often diagnosed with the aggressive kind of this disease, and it may be a very different clinical entity (Salinas, Tsodikov et al. 2014). However, much more work needs to be done with regards to this risk association to affirm this.
5.4.4.2. BMI, smoking tobacco, and alcohol consumption (external factors) at recruitment and risk of prostate cancer:

In our previous approach, we combined the effect of the three external factors to extract as much from the prevalent factors common among New Zealand men and risk of PCa and not miss any SNP of interest. However, in this current analysis, we split the three parameters, and analyzed the effect they have individually and with age as well as risk for PCa with statistical adjustments.

The data for the demographic analyses related to high BMI, tobacco smoking, and alcohol consumption has previously been reported (Vaidyanathan, Naidu et al. 2017).

5.4.4.3. Gene-environment interaction and risk of prostate cancer and effect of adjustment for age:

Knowledge of gene x environment interaction is important for risk prediction and the identification of certain high-risk populations to inform public health strategies for targeted prevention (Rudolph, Chang-Claude et al. 2016). We associated the environmental factors with the genotypes of the men in our study to identify the risk alleles for specific kind of external factors such as BMI, smoking tobacco and alcohol consumption. Since these factors play an important role in the risk association of PCa and yet can be controlled by individuals, it is therefore of importance to understand and limit this disease.

5.4.4.3.a. SNP genotyping, the effect of environmental factors, and of age as a risk of aggressive prostate cancer vs healthy controls:

We had previously identified 14 SNPs when we analyzed the data for gene x environment interactions without any adjustments (Table 5.12) (Vaidyanathan, Naidu et al. 2017). This gave us a good idea of the influence of environmental factors on various SNPs in and near certain genes, and the prevalent environmental conditions in New Zealand. Of the 14 SNPs, three were found near the gene MYEOV (Myeloma Overexpressed)- rs7931342, rs10896438,
rs11228565; two near the gene KLK3 (Kallikrein-3)- rs2659122, rs17632542; and one each near the genes MSMB (Microseminoprotein Beta)- rs7920517, FADS2 (Fatty acid desaturase 2)- rs2727270, LEP (Leptin)- rs10244329, PPAR-γ (Peroxisome Proliferator-Activated Receptor Gamma)- rs17793693, CCHCR1 (Coiled-Coil alpha-Helical Rod protein1)- rs130067, AKR1C3 (Aldo-Keto Reductase family 1 member C3)- rs12529, SLC26A6 (Solute carrier family 26 member 6)- rs887391, and NUDT11 (Nucleoside Diphosphate-linked Moiety X Motif 11)- rs5945619; and in the region 8q24- rs6983561.

These results were partly expected and partly novel to New Zealand conditions and the risk of aggressive PCa. MYEOV is a putative oncogene (Szyfter, Wierzbicka et al. 2016), and it made absolute sense that the highest number of SNPs were recorded in this gene with regards aggressive PCa in our population (Vaidyanathan, Naidu et al. 2017). The genes KLK3, and MSMB are both involved in the androgen metabolism pathway were understandably identified as statistically significant in our study, due to their proven risk association to PCa, and same with the SNP in AKR1C3 (Karunasinghe, Han et al. 2012; Karunasinghe, Lange et al. 2013; Karunasinghe N 2014; Vaidyanathan, Naidu et al. 2017) and the SNP in CCHCR1, which has been previously reported in rheumatoid arthritis- a possible side-effect of androgen deprivation therapy for PCa (Orozco, Goh et al. 2013). The gene SLC26A6 is a fusion gene and plays a vital role in the development and progression of a number of cancers and is interestingly just 10Mb centromeric to the gene KLK3, which we have already identified as an important gene of interest with regards studies on PCa (Lambros, Wilkerson et al. 2011). NUDT11 is a paralogous human gene, and is predominantly expressed in the testes, and assumed to be playing a major role in signal transduction (Hidaka, Caffrey et al. 2002; Hua, Hidaka et al. 2003). Various GWAS and case control studies have also indicated about the susceptibility locus at NUDT11 being involved with the risk of PCa (Eeles, Kote-Jarai et al. 2008; Camp, Farnham et al. 2009; Fitzgerald,
Kwon et al. 2009). The presence of a SNP as risk for PCa in the gene desert region of 8q24 has also been observed in a number of cancers including the prostate (Wasserman, Aneas et al. 2010).

With no direct connection yet established between obesity and risk of PCa, it was interesting to find SNPs associated with risk of PCa in our population in 3 genes. The genes FADS2, LEP, PPAR-γ are associated with obesity and diabetes mellitus which is a major risk of PCa (Freedland and Aronson 2004; Parikesit, Mochtar et al. 2016; Vaidyanathan, Naidu et al. 2017). This is interesting because New Zealand has the third highest adult obesity rate among Organisation for Economic Co-operation and Development countries (2015), and is a major external factor in the potential risk for aggressive PCa (Vidal, Howard et al. 2014).

When we, next, adjusted the SNP genotyping data for age of the cohort and continued to analyse the data, we found certain SNPs to have lost their power of statistical significance on risk of aggressive PCa, and certain SNPs were identified statistically significant which were not identified without the adjustment. SNPs rs632148 and rs6502051 in genes SRD5A2 (Steroid 5α-reductase type 2) and FASN (Fatty Acid Synthase) respectively were identified as statistically significant to the risk of aggressive PCa when compared to healthy controls. The gene SRD5A2 has previously been reported by groups working on various aspects related to and causing PCa in Caucasian populations and not restricted only to studies discussing its role in the quality of sperms (Zhao, Wu et al. 2012). It is well established that with progressing age, there is a drop in testicular function, and thus certain genes pertaining to virility, including SRD5A2, may be functioning differentially (Perheentupa and Huhtaniemi 2009). The SNP in a gene pertaining to obesity (Nguyen, Ma et al. 2010; Chavarro, Kenfield et al. 2013)- FASN also identified as a risk for aggressive PCa is also in line with the theory that ageing may cause certain physiological alterations leading to major effects such as, and not limited to, PCa (Vaidyanathan, Karunasinghe et al. 2016). Since obesity is classically
considered to be proportional to progressing age (Han, Tajar et al. 2011), we feel that our findings are further strengthening the theory of age as a risk factor for PCa (Vaidyanathan, Karunasinghe et al. 2016), especially aggressive PCa. The other SNPs that were identified to be statistically associated as risk for aggressive PCa, even after the adjustment for age, were rs7931342, rs10896438, and rs11228565 near the gene\textit{MYEOV}; rs7920517 near the gene\textit{MSMB}, rs2659122 near the gene\textit{KLK3}; rs10244329 near the gene\textit{LEP}; rs130067\textit{CCHCR1}; and rs887391\textit{SLC26A6}.

Next, we adjusted the data for BMI, and identified that apart from the SNP rs6502051 near the gene\textit{FASN}, the other SNPs that were identified to have statistical significant association as risk for aggressive PCa when compared to healthy controls after adjusting for age remained significant. This helps us define the role of BMI as risk for aggressive PCa with ageing (Han, Tajar et al. 2011).

We then adjusted the data for BMI and age. Interestingly, instead of getting a lesser number of SNPs associated with the risk of aggressive PCa, we identified three more SNPs. Since the data was adjusted for BMI and age, this, statistically, implies the effect of alcohol consumption and tobacco smoking on our health. The additional SNPs identified as significantly associated with the risk of aggressive PCa were rs3918256, rs5945619, and rs6502051 present near the genes\textit{MMP9 (Matrix metallopeptidase 9), NUDT11, and FASN} respectively. The SNPs in gene\textit{FASN} has previously been discussed with regards its role as risk for aggressive PCa, but the SNP in the gene\textit{MMP9}- an inflammation marker (Bruschi, Bianchi et al. 2014) was not previously identified when seeing the role gene x environment interaction plays. Both, tobacco smoking and alcohol consumption have been studied in the recent past to be altering the levels of expression of MMP9 protein (Koken, Gursoy et al. 2010; Watson, Benton et al. 2010).
Next, we adjusted the data for tobacco smoking only, in order to identify the risk age, BMI, and alcohol consumption have as a risk of aggressive PCa when compared to healthy controls. We identified two new SNPs, compared to the result generated by adjusting the data for age, being rs12529, in the gene *AKR1C3* and rs799923 near the gene *BRCA1*. The crosstalk between tobacco smoking and the SNP rs12529 in the gene *AKR1C3* has previously been explored by our group (Karunasinghe, Zhu et al. 2016). Interestingly, the identification of the SNP rs799923 near the gene *BRCA1*, a tumour suppressor (Silver and Livingston 2012), indicates that with progressing age, certain genes may function differently in the presence of external stresses such as alcohol consumption (McDonald, Goyal et al. 2013).

We got further evidential proof with regards the effect of age on the expression and effect of tumour suppressor genes such as *BRCA1* on diseases such as aggressive PCa, when we analyzed the data after adjusting for tobacco smoking and age and found that the gene was no longer significantly associated as a risk for the disease. Interestingly the significant association of risk of aggressive PCa was lost in the SNPs in the genes *AKR1C3* and *KLK3* too. The result pertaining to the SNP in the gene *AKR1C3* is interesting. As aforementioned, we have found some interesting correlations between the gene *AKR1C3*, tobacco smoking and the risk of PCa (Karunasinghe, Zhu et al. 2016) and when we adjusted for age, the role of the SNP as a potential risk for aggressive PCa, compared to healthy controls, was not found to be statistically significant. We believe age-long smoking tobacco has a more potent effect on the risk of aggressive PCa rather than not. Consistent with the effect of adjusting the data for BMI and age, we identified SNP 632148 in the gene *SRD5A2* to be significantly associated with the risk of aggressive PCa. This, we believe, helps understand the nexus between ageing and the effect of certain genes and the influence of external factors leading to oxidative stress in a body.
In the final set of adjustments of our data to analyse the effect of SNPs as risk of aggressive PCa, we considered alcohol consumption and the combination of alcohol consumption and age. Interestingly, the SNP rs1799977 present in the gene \textit{MLH1} (\textit{MutL homolog 1}), which plays a major role in DNA (deoxyribonucleic acid) mismatch repair (Pal, Permuth-Wey et al. 2008), and more so because rs1799977 is an exonic SNP (2017; Vaidyanathan, Naidu et al. 2017). DNA mismatch repair mechanism is an important fightback against major diseases such as cancer (Hsieh and Yamane 2008). SNPs in the genes \textit{SEP15} and \textit{FASN} are found significantly associated with risk of aggressive PCa when compared with healthy controls with adjustments for just alcohol and combination of alcohol and age respectively. The effects of smoking and BMI have always been a matter of controversy, but according to Kaufman et al., (2012), tobacco smoking can have a wide range of effects including limited physical activities, and it itself being a “gateway” habit, the effect on increasing BMI and obesity should be accepted (Kaufman, Augustson et al. 2012).

The use of such combinations to adjust the data and extract the fine points of a case-control study is quite a unique approach on its own, however, the SNPs in the various genes that we have identified as a risk of aggressive PCa when compared to healthy controls is quite interesting. With as many as five SNPs across three genes- \textit{MYEOV}, \textit{MSMB}, and \textit{SLC26A6} that remained significantly associated as risk for aggressive PCa, it is beyond doubt that these are the most important genes of interest with regards to similar studies. Having said this, it is worthy of bringing to notice that studies in larger populations need to be done to validate these results, though (Figure 5.9).
5.4.4.3.b. **SNP genotyping, the effect of environmental factors, and of age as a risk of aggressive prostate cancer vs non-aggressive prostate cancer:**

A similar approach was employed to determine the SNPs in genes of interest with regards the risk of aggressive PCa when compared to non-aggressive PCa. If the logic of progression of PCa holds true, non-aggressive PCa is the most crucial stage, as due to cell division with accumulation of cancer cells, and a prolonged weakening of immune cells, non-aggressive PCa could progress to aggressive PCa (Vaidyanathan, Karunasinghe et al. 2016; Vaidyanathan, Naidu et al. 2017). We believe that this is one of the most important sets of data that we have analyzed thus far, as knowledge of these SNPs and corresponding genes is important to arrest non-aggressive PCa from progressing to aggressive PCa.

We first analyzed the data without adjustment for any of the four afore mentioned factors, for the gene x environment effect as a risk of aggressive PCa compared to non-aggressive PCa and has been explained in details in one of our recent publications (Vaidyanathan, Naidu et al. 2017). One SNP each in the genes *SRD5A2*- rs632148, *MLPH* (*Melanophilin*)-rs2292884, *PODXL* (*Podocalyxin-like*)- rs3735035, *LEP* (*Leptin*)- rs10244329, *TLR4* (*Toll-like receptor 4*)- rs11536889, *SLC26A6*- rs887391, *KLK3*- rs17632542, and *MMP9*-

![](image.png)

Figure 5.9 Various pathways and the genes identified to be significantly associated with a risk of aggressive prostate cancer (compared to healthy controls).
rs3918256 were identified as statistically significant risk of aggressive PCa (compared to non-aggressive PCa). As expected, we identified that there is a general trend of a typical textbook-like analysis of progression of any cancer. We identified SNPs in a fusion gene-

SLC26A6 which is well established to aid the development of human cancers (Lambros, Wilkerson et al. 2011; Vaidyanathan, Naidu et al. 2017); MMP9 and TLR4- genes involved in the inflammation pathway (Zhao, Zhang et al. 2014; Vaidyanathan, Naidu et al. 2017); PODXL- a gene encoding for the cell-adhesion glycoprotein which has previously been reported to be associated with aggressive tumour phenotype and poor prognosis in various cancers (Larsson, Johansson et al. 2011; Boman, Larsson et al. 2013; Vaidyanathan, Naidu et al. 2017); along with genes pertaining to steroid levels- SRD5A2, and overexpressed in the estrogen receptor - MLPH (Casey, Neville et al. 2006); along with a gene pertaining to obesity- an import external risk factor for aggressive PCa (Vaidyanathan, Naidu et al. 2017) and KLK3- involved in the androgen metabolism pathway (Vaidyanathan, Naidu et al. 2017). The data is indicative of a strong gene x environment interaction leading to the progression of the disease.

We then adjusted the data for age to identify the genes which may be influenced by progressing age (Vaidyanathan, Karunasinghe et al. 2016). Interestingly, only four of the aforementioned eight SNPs remained significantly associated with the risk of aggressive PCa when compared to non-aggressive PCa. These were identified as the SNPs in the genes SRD5A2, PODXL, LEP and MMP9. Incidentally, only these four SNPs remained significantly associated as risk for aggressive PCa when compared with non-aggressive PCa across all statistical adjustments.

The role between inflammation and the development of cancer is a very well established nexus (Coussens and Werb 2002; Rakoff-Nahoum 2006). With the progression of cancer, the tissue(s) may change drastically, which may trigger certain homeostatic processes of tissue
repair, and the recruitment of inflammatory leukocytes (Rakoff-Nahoum 2006) and affect innate immunity as well. Not only MMP9, but other members of this family of enzymes with their role in the evolution of the immune system are well known to regulate certain inflammatory and repair processes and hence may be used for predictory analysis for various cancers (Parks, Wilson et al. 2004). The fact that a SNP in this gene was identified as significantly associated as risk of aggressive PCa is understandable.

PODXL is cell-adhesion glycoprotein which is also associated with a number of aggressive tumour outcomes (Taniuchi, Furihata et al. 2016). This transmembrane glycoprotein is expressed in a number of cancers including ovarian (Cipollone, Graves et al. 2012), epithelium (Meng, Ezzati et al. 2011) and prostate (Vaidyanathan, Naidu et al. 2017). PODXL causes an increase in cell migration as well as invasion, leading to an increase in the MMP expression (Taniuchi, Furihata et al. 2016), which has an established role in inflammation (Rakoff-Nahoum 2006) and innate immunity.

One of the other important genes that upregulates the function of some members of the MMP family (Fan, Gan et al. 2015), and is significantly associated with obesity and the risk of a number of cancers is LEP (Parekh, Chandran et al. 2012). There have been a number of studies to define the role of obesity in carcinogenesis (Dutta, Ghosh et al. 2012), but it is usually poorly understood (Parekh, Chandran et al. 2012). With an increase in the world population’s BMI, it is vital to identify means to understand the progression of various diseases, including aggressive PCa owing to the SNPs and thereby altered expression of obesity-related genes such as LEP.

As expected, the SNP rs632148 present near the gene SRD5A2 was identified to be significantly associated with the risk of aggressive PCa when compared with non-aggressive PCa, just as was when compared to the healthy controls. The enzyme produced by the gene SRD5A2 is important for the development and growth of the prostate gland (Ge, Wang et al.
and assists in the conversion of the male sex hormone, testosterone into the more effective androgen dihydrotestosterone (Rajfer 2000). With testosterone-levels being a matter of debate amongst urologists with regards the risk of PCa (Klap, Schmid et al. 2015), it is interesting to find $SRD5A2$ as significantly associated with risk of aggressive PCa in our population, because New Zealand is predominantly an overweight population (2015), and increase in BMI reduces testosterone levels (Fui, Dupuis et al. 2014). This reduction in testosterone levels with increased BMI is interesting, as we feel, an increase in BMI, may increase the dilution factor due to an increase in the overall size of the body, but further work needs to be done to prove this.

The New Zealand story (gene x environment interactions and risk of aggressive PCa) gets firmly knit when we put the results in this section together (Figure 5.10). It is well established that obesity has a major contribution in the inflammatory pathway (Monteiro and Azevedo 2010), which in turn leads to the progression of cancers into advance stages (Coussens and Werb 2002; Rakoff-Nahoum 2006). Moreover, age and obesity have a role leading to alterations in testosterone levels, as previously discussed (Vaidyanathan, Karunasinghe et al. 2016), and this hormonal imbalance, in turn, is a risk for aggressive PCa (Karunasinghe, Lange et al. 2013; Klap, Schmid et al. 2015). Thus, the effect of age on and with obesity may be playing a major role in our population with regards the total number of cases with aggressive PCa. This, we believe, is a very unique finding.
5.4.4.3.c. **SNP genotyping, the effect of environmental factors, and of age as a risk of non-aggressive prostate cancer vs healthy controls:**

Finally, we analyzed the data with and without various statistical adjustments to understand the initiation of PCa in our population and effect of age by comparing non-aggressive PCa with healthy controls. We identified only four genes with one SNP in and/or near it that was identified as statistically significant with the risk of non-aggressive PCa. They being rs2292884 in the gene *MLPH*, rs3735035 in the gene *PODXL*, rs11536889 in the gene *TLR4*, and rs4965373 near the gene *SEPS1* (*Selenoprotein 1*). With 3 out of 8 genes identified to be common with the risk of aggressive PCa without any statistical adjustments, it indicates that there is a continuation with regards the alteration of certain gene functions with the schematic progression of the disease. Interestingly, however, none of the SNPs were identified to bear any significant association with the risk of non-aggressive PCa after various statistical adjustments including for age were performed. This implies that perhaps the gene x environment interactions, rather the genes on their own play the most important role in the initiation of diseases such as PCa.

The fact that a single gene involved with selenium metabolism- *SEPS1* was also significantly associated with the risk of non-aggressive PCa cannot be ignored, as yet another
selenoprotein- *SEP15* was associated with risk of aggressive PCa (compared to healthy controls) when statistically adjusted for certain demographic parameters, as discussed above. The deficiency of trace elements such as selenium in the New Zealand soil is a well-established fact (Hewitt and Dymond 2013), and in the absence of the same, certain people take dietary supplements. However, a direct correlation between the role played by these dietary supplements and risk of PCa was recently identified (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2013; Karunasinghe, Zhu et al. 2016). Two of the other three genes involved are pertaining to the inflammatory pathway- *TLR4* and *PODXL*, which again can be due to the side-effect of the prevalence of high number of tobacco smokers in New Zealand (2015), and the third one is overexpressed in the estrogen receptor- *MLPH*, which may be influenced by the low levels of Vitamin D among our cohort because of the lesser exposure to sunlight due to ageing (Howlader N; Ministry of Health 2012) (Table 5.15).

Therefore, it does seem that the inflammatory pathway is one of the most important pathways for the initiation of PCa, along with the local factors such as life-long consumption of food low in selenium, and exposure to low levels of Vitamin D due to various factors with progressing age, and with the effect of hormones pertaining to specific organ of interest that eventually may be critical. The gene x environment interaction with the adjustment for age has brought a completely new way of looking at and understanding the risk for aggressive PCa based on the data generated from our cohort.
Table 5.15 New Zealand factors” and risk of non-aggressive prostate cancer

<table>
<thead>
<tr>
<th>New Zealand factor(s)</th>
<th>Reference</th>
<th>Gene involved</th>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Selenium levels in soil (leading to lower dietary intake)</td>
<td>(Hewitt and Dymond 2013)</td>
<td>SEPS1</td>
<td>rs4965373</td>
</tr>
<tr>
<td>Low sun exposure (leading to low Vitamin D levels)</td>
<td>(Ministry of Health 2012)</td>
<td>MLPH</td>
<td>rs2292884</td>
</tr>
<tr>
<td>High tobacco smoking (leading to inflammation)</td>
<td>(2015)</td>
<td>PODXL</td>
<td>rs375035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TLR4</td>
<td>rs1153689</td>
</tr>
</tbody>
</table>

5.4.5. Conclusions

SNPs, being the most commonly observed variations in the genome, are ideal candidates for identification of biomarkers for various diseases (Vaidyanathan, Naidu et al. 2017). Genotyping SNPs and observing the gene x environment interactions is a very useful tool to identify the various local factors and their effect on genes leading on to a bottle-neck population with a particular condition- in this aggressive PCa.

We have identified a number of important individual lifestyle factors and their effect (either due to lifestyle exposure or due to ageing) as risk factors for PCa and aggressive PCa. We propose that the inflammatory pathway is one of the most important pathways responsible for initiating the disease, and certain local demographic factors such as obesity and tobacco smoking play crucial roles in driving non-aggressive PCa to the aggressive stage. SNPs in a putative oncogene (MYEOV) play a very influential role as risk for aggressive PCa. These findings are crucial for planning larger scale studies, because, although we recruited men of European ethnicity in our study, and genotyped SNPs that were identified as significantly associated as risk for PCa in various European populations, we could define a
clear dependence of age in the progression of the disease based on gene x environment aspects. We propose that further studies based on our case-control analyses should be carried out to define specific biomarkers on a regional-basis, as this will help develop better diagnostic and treatment methods which will be tailor-made.
5.5. SNP-SNP interactions as risk factors for aggressive prostate cancer

Abstract: Prostate cancer (PCa) is one of the most significant male health concerns worldwide. Single nucleotide polymorphisms (SNPs) are becoming increasingly strong candidate biomarkers for identifying susceptibility to PCa. We identified a number of SNPs reported in genome-wide association analyses (GWAS) as risk factors for aggressive PCa in various European populations, and then defined SNP-SNP interactions, using PLINK software, with nucleic acid samples from a New Zealand cohort. We used this approach to find a gene x environment marker for aggressive PCa, as although statistically gene x environment interactions can be adjusted for, it is highly impossible in practicality, and thus must be incorporated in the search for a reliable biomarker for PCa. We found two intronic SNPs statistically significantly interacting with each other as a risk for aggressive prostate cancer on being compared to healthy controls in a New Zealand population.

5.5.1. Introduction

Prostate cancer (PCa) is highly prevalent, and around 1 in 6 patients are at risk of developing the aggressive form of the disease (Cooperberg, Vickers et al. 2010). It has become one of the most significant male health concerns worldwide (Jemal, Bray et al. 2011). An individual is diagnosed as having high-risk or aggressive PCa based on the classification by the American Urological Association (Thompson, Thrasher et al. 2007), when the clinical T stage $\geq$T2c, and/or the Gleason score $\geq$8, and/or the serum prostate serum antigen (PSA) level $>20$ng/ml (D'Amico, Whittington et al. 1998).

Although a hereditary aspect is well known for this disease (Bratt 2002), various studies have also shown that genetic interactions with biological and behavioral factors play an important role in the overall risk and prognosis of PCa (Schaid 2004; Karunasinghe, Han et
al. 2012; Karunasinghe, Lange et al. 2013). Variations in the genome are a major contributor to the differences in disease susceptibility amongst individuals (Tweardy and Belmont 2009). Single nucleotide polymorphisms (SNPs) are the most commonly identified variations in a genome.

Analysing the role of SNP-SNP interactions and epistasis (Cordell 2002) is very appealing among researchers working on risk factors for various cancers (Hartwig 2013; Su, Yao Shugart et al. 2013; Jamshidi, Fagerholm et al. 2015), including prostate cancer (Tao, Feng et al. 2012). Here we have identified a SNP-SNP interaction as a risk factor for aggressive PCa, by comparing the data generated after carrying out SNP genotyping using the SEQUENOM MassARRAY iPLEX® assay, and the TaqMan® assay (depending on the gene of interest) from the DNA extracted from blood samples. These samples were taken from a New Zealand cohort of men with self-reported European ethnicity that have been clinically diagnosed with aggressive and non-aggressive PCa, and healthy controls with no reported symptoms of the disease. Symptoms include increased urination during night time along with a frequent urge to urinate problems maintaining a steady flow of urine, hematuria and dysuria (Hodgson, Obertova et al. 2012). Our results indicate a strong influence of gene x environment interaction in overall gene expression and epistasis.

5.5.2. Methods

5.5.2.1. Study population

The study cohort is same as in section 5.1.2.1.

5.5.2.2. Collection and processing of blood samples

The participants blood sample collection and processing has been explained in Section 5.1.2.2.
5.5.2.3. Selection and genotyping of SNPs

The selection and genotyping of SNPs has already been explained in Section 5.1.2.4 of this thesis.

5.5.2.4. Statistical analysis

26 SNPs were removed for being in linkage, and a further 5 SNPs were removed for failing the Hardy-Weinberg Equilibrium (HWE) in the healthy controls, thereby reducing the total SNPs analyzed to 105 (colour coded in Dataset 1). SNPs that failed the HWE in patients with PCa were still considered for analysis, as the SNPs may have failed to be in equilibrium in the patient population due to the influence of the risk allele and hence should not be ignored from a case-control study like ours (Namipashaki, Razaghi-Moghadam et al.; Yong Zou and Donner 2006). The statistical significance was set to be \( p \leq 0.0001 \) (Purcell, Neale et al. 2007).

Analysis of the data for SNP-SNP interactions associated with aggressive PCa was carried out using PLINK software version 1.07 (Purcell, Neale et al. 2007; Vaidyanathan, Naidu et al. 2017). PLINK's clustering approach is based on the genome-wide average proportion of alleles shared identical-by-state (IBS) between two individuals SNPs, i.e., pairing up the SNPs based on similarity of genetic identity (Purcell, Neale et al. 2007). This IBS clustering is used in order to test if the SNPs of two individuals belong to the same population or not (Vaidyanathan, Naidu et al. 2017). Following this stratification, we performed a standard case-control association test using the Cochran-Mantel-Haenszel test (1 degree of freedom) to analyse the SNP-disease association that is conditional on the clustering (Vaidyanathan, Naidu et al. 2017). The slower ‘--epistasis’ command was used to test for epistasis using logistic regression (Purcell, Neale et al. 2007). It is the most accurate test to define SNP-SNP interactions using PLINK (Purcell, Neale et al. 2007).
5.5.3. Results

Table 5.16 shows the statistically significant SNP-SNP interactions discovered in patients with aggressive PCa when compared to healthy controls. The results obtained for other categorical analyses are not discussed here, as they were not statistically significant in our study and have been mentioned in Supplementary Table 2 (available online). The SNP rs2121875, an intronic SNP present in chromosomal position 5p12 near the fibroblast growth factor 10 (FGF10) gene (Kote-Jarai, Olama et al. 2011), has been identified to be associated with the SNP rs4809960, an intronic SNP present in chromosomal position 20q13 near the gene cytochrome P450 family 24 subfamily A member 1 (CYP24A1) (Holt, Kwon et al. 2010), such that the latter SNP raises the odds of having the prior.

Table 5.16 Statistically significant SNP-SNP interactions discovered in patients with aggressive PCa when compared to healthy controls.

<table>
<thead>
<tr>
<th>CHR1</th>
<th>SNP1</th>
<th>CHR2</th>
<th>SNP2</th>
<th>OR_INT</th>
<th>STAT</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>5p12</td>
<td>rs2121875</td>
<td>20q13</td>
<td>rs4809960</td>
<td>2.918</td>
<td>15.77</td>
<td>7.15E-05</td>
</tr>
</tbody>
</table>

Table legends: CHR1: chromosome of first SNP, SNP1: Identifier for first SNP, CHR2: Chromosome of second SNP, SNP2: Identifier for second SNP, OR_INT: Odds ratio for interaction, STAT: Chi-square statistic 1df, p: Asymptotic p-value

5.5.4. Discussion

Epistatic effects that are crucial to define various biologically-intuitive models of interaction between two SNPs have already been observed in a variety of species (Hartwig 2013). We believe this is the first study on SNP-SNP interactions associated with aggressive PCa carried out with patients from a New Zealand population.

The SNP rs4809960 in the gene CYP24A1 has been reported by Holt et al., (2010) to be associated with prostate cancer-specific mortality, and was not evolutionarily conserved (Holt, Kwon et al. 2010). It was also found to have an effect on the body mass index (BMI), but due to a small sample size the hazard ratios for the BMI strata were not considered reliable enough to be reported (Holt, Kwon et al. 2010). The protein encoded by CYP24A1
initiates the degradation of the physiologically active form of Vitamin D3 (VD3) (Di Rosa, Malaguarnera et al. 2011). VD3 is an important hormone that is actively involved in regulating cell proliferation in the prostate, and has also been identified to have increased expression in PCa cell lines (Lou, Qiao et al. 2004). It is well established that, with ageing, the skin cannot synthesize VD3 as effectively as desirable and the kidney’s ability to convert VD3 to its active form decreases (Nair and Maseeh 2012). This is of relevance because PCa has always been considered as a disease of elderly men (Nelen 2007) who have had less exposure to sunlight and thereby Vitamin D3 (2012). It is even more intriguing for the other epistatic SNP to be identified in FGF10.

According to Paul et al. (2013), during mesenchymal development, FGF10 protein can trigger PCa development through increased androgen receptor expression in the neoplastic epithelium (Corn, Wang et al. 2013). It is also worthy to mention that FGF10 is closest to FGF7 based on its evolutionary history (Emoto, Tagashira et al. 1997), and according to Emoto et al. (1997), is suggested to have no activity for fibroblasts (Emoto, Tagashira et al. 1997). We do not agree with this, because fibroblasts in certain organs, senesce due to aging (Campisi 1998), and can promote tumour invasion (Coppe, Desprez et al. 2010). This logical progression of ageing-led senescence and promotion of tumour invasion holds true for ageing and risk of aggressive PCa (Vaidyanathan, Karunasinghe et al. 2016) as well, and is in line with the effect of ageing on VD3 encoded by the gene CYP24A1, due to epistasis as identified by us, rather than the protein FGF10 not having a role in senescence.

We suggest that the intronic SNP rs2121875 in the gene FGF10 may be causing alterations in gene expression, perhaps due to the prevalent external/ environmental conditions in the elderly men with PCa. Our theory is based on the recent discovery in a study by Zhang et al. (2007) that even intronic SNPs (such as the ones identified in FGF10 and CYP24A1) can change the outcome and usage of exons (Zhang, Bertolino et al. 2007; Tazi,
Bakkour et al. 2009). This unique and novel epistatic finding emphasizes the fact that intronic SNPs (and SNP-SNP interactions) can also have a significant effect on the risk of diseases such as aggressive PCa, and need to be investigated further.

**Data availability**

The cleaned raw data after removing the personal information of the participants such as name, address, date of birth, name of general practitioner and/or clinic, and etc. is provided in the form of supplementary material (available online).

Supplementary Table 1: Raw data for the current study.

Supplementary Table 2: Total results for various analyses for this research report. The results that were significant ($p \leq 0.0001$), was first generated, and then the second part of the output was carried out, which was useful in identifying the significant epistatic result (in this case only in Aggressive PCa vs healthy controls).
5.6. Chapter Summary

Thus, owing to the various lifestyle factors which were identified to be playing an important role in the risk of aggressive PCa in our cohort (Section 5.1), we looked into the effect of age and the role of each behavioral parameter individually and in conjunction with age-related statistical correction. We first proposed the effect of ageing and the risk of PCa (Section 3), and then went ahead with the analyses based on the data obtained from our cohort. This aided us to identify the SNPs which we could have potentially missed to identify and thereby a major flavor to a typical case-control study would have been lost.

The gene x environment interaction and the influence of various factors based on progressing age identified in our cohort (Section 5.4) undeniably brought forth certain interesting results. These findings brought us closer to the problem “in home” and helped us tighten our story further. We believe that these results should further be utilized to assess data from different New Zealand cohorts. The SNP-SNP interaction analysis (Section 5.5) also yielded a unique and thus far unreported epistatic relation between two intronic SNPs in different genes, along with the underlying effect of ageing and the risk of PCa.

This multi-directional approach using PLINK software helped us identify and gather knowledge about several SNPs- some known and well-established, but some unique too and seemingly specific to the New Zealand scenario. This motivated us to ponder further and try and identify if we missed identifying any risk SNPs using artificial intelligence (chapter 6.1)- a major leap for us from human genetics to employing advanced bioinformatics tools to define the risk for aggressive PCa.
"Artificial Intelligence will never be a match for natural stupidity."

- Joseph Addison
Chapter 6. Artificial Intelligence- Artificial Neural Networks

Chapter preamble:

Although the long-term goal of the neural-network community remains the design of autonomous machine intelligence, the main modern application of artificial neural networks is in the field of pattern recognition. In the sub-field of data classification, neural-network methods have been found to be useful alternatives to statistical techniques such as those which involve regression analysis or probability density estimation.

The most widely applied neural network algorithm in image classification remains the feedforward backpropagation algorithm. This chapter is devoted to explaining the basic nature of this classification routine and its application to identify the SNPs and risk of aggressive prostate cancer.

The chapter herewith attached (chapter 6.1) has been written in lines of a journal article. Although very novel, we believe that this approach has not yet been established using different samples across various populations and or diseases of interest, and much more work needs to be done before it can be patented.
6.1. Artificial Neural Networks: Unravelling “hidden” SNP biomarkers for aggressive prostate cancer- a New Zealand case-control study

Abstract: Artificial Neural Networks is a crucial machine-learning method employed to define the role of advanced bioinformatics and the identification of certain huge unprocessable data for human brain. This method of creaming the undefined role of certain factors on the risk of diseases, such as aggressive prostate cancer is not only unique in that the machine-learning tools are “self-trained” and run with minimal human intervention(s). we have compared the results, herewith, between the association of certain SNPs are risk for aggressive PCa obtained by using established software and with the one developed by artificial intelligence, and have narrowed down our overall number of SNPs associated with risk of aggressive disease considerably.

6.1.1. Introduction

Prostate cancer (PCa) is one of the most significant non-skin cancer male health concerns worldwide (Vaidyanathan, Naidu et al. 2017). With the alarming estimate that at least 1 in 6 PCa patients is vulnerable to risk of aggressive PCa (Cooperberg, Vickers et al. 2010), the identification of a predictive biomarker of this disease is of much importance. More so for New Zealand, because the highest recorded rate of men with PCa relative to the population of men is observed in this region (Jemal, Bray et al. 2011; Karunasinghe, Han et al. 2012) along with a high number of tobacco smokers in New Zealand (2015), deficiency of trace elements such as selenium in the New Zealand soil (Hewitt and Dymond 2013), and New Zealand being an overweight population (2015).

With certain external, biological and behavioral factors established as playing crucial role in the overall risk and prognosis of PCa (Schaid 2004; Karunasinghe, Han et al. 2012;
Karunasinghe, Lange et al. 2013; Vaidyanathan, Naidu et al. 2017), SNPs, whatever the rationale in selection for various studies are increasingly appealing biomarker candidates for the identification of PCa susceptibility (Perkel 2008; Tao, Wang et al. 2012; Goh, Saunders et al. 2013; Van den Broeck, Joniau et al. 2014). The function of these SNPs, unfortunately, remains largely unknown and data describing their correlation with clinical factors and/or their interplay with other genetic and non-genetic factors are rare- mainly due to the large sample sizes that is needed for sufficient statistical power (Lindstrom, Schumacher et al. 2011) to prove the theories, and unavailability of software to handle large sets of data adjusting for various environmental and epigenetic factors without missing the “hidden” truth.

Machine-learning is a section of artificial intelligence that uses a wide range of statistical, probabilistic and optimization techniques that permits computers to “learn” and define “hidden” hard-to-discern patterns from large, noisy and/or complex data sets (Cruz and Wishart 2006). To enhance the screening method and produce accurate results in the detection of certain cancers, the artificial neural networks (ANN) is used to produce accurate results (Agatonovic-Kustrin and Beresford 2000; Stephan, Cammann et al. 2002; Adetiba and Olugbara 2015; Devi, Ravi et al. 2016). ANN is a biologically motivated mathematical or computational model produced from hundreds of individual units, or, artificial neurons, associated with coefficients (weights) which represent the neural structure (Cruz and Wishart 2006). This neural structure is also referred to as processing elements (PE) as they process data. Every PE has weighted inputs, output data, and transfer function. PE is typically an equation which corresponds both inputs and outputs.

Different groups have tried to tackle the problem of diagnosing the risk of PCa using ANN in various ways (Hu, Cammann et al. 2013). Saritas et al., (2010), for example involved the parameter of age and the results of the free PSA levels and risk of PCa in their ANN analysis.
(Saritas, Ozkan et al. 2010). Snow, *et al.*, (1994) carried out one of the first pilot studies to predict biopsy results, for example, using the datasets of men who had undergone radical prostatectomy (Snow, Smith et al. 1994). However, having said so, it is fair to say that the approach we have taken is very novel because, not only did we consider quite a number of SNPs genotype data, but also, we are looking into aggressive PCa specifically.

Broadly speaking, there are three layers used in the architecture of ANN- the input layer, the hidden layer and the output layer. The number of each layer depends on the input images fed which connects to each layer with the neurons. The Figure 6.1 illustrates the three layers connectivity with neuron structure where nodes are used for each layer, thereby identifying the risk associations which were previously unidentifiable, and the Figure 6.2 is an overall flow chart of the procedure followed.

![Figure 6.1: Three layers connectivity with neuron structure.](image)
Herewith, we are presenting the outcome of ANN on the data generated using SNP genotyping for 136 SNPs, located in 66 genes and some undefined chromosomal locations among a population of 254 men with PCa (197 with clinically diagnosed aggressive PCa and 57 with clinically diagnosed non-aggressive PCa) and 369 healthy controls from New Zealand. We have compared the SNPs identified as significant risk for PCa as well as for aggressive PCa using an IBS clustering method- which is used for such data and by using ANN– which is soon making a mark for itself in biomedical research. This is a very unique study and the first of its kind in a New Zealand cohort.
6.1.2. Methods

6.1.2.1. Selection of Data Set

The PubMed database was screened for research articles using key terms “prostate cancer” and/or “aggressive prostate cancer”, along with “SNP genotyping sequenom massarray iplex”, and/or “genome-wide association studies” to identify a list of target SNPs to be worked on, as reported in Vaidyanathan et al., 2017 (Vaidyanathan, Naidu et al. 2017).

Genome wide association analysis (GWAA) was used to provide a comprehensive evaluation of multiple genes with polymorphisms that interact in the same pathway/s, in line with proposals made by Hsing et al., (2008) (Hsing, Chu et al. 2008) and Kwon et al., (2012) (Kwon, Holt et al. 2012), who claim that studies on GWAA are a better way to study multifactorial diseases such as PCa.

6.1.2.2. Data tidying and pre-processing

A total of 39 SNPs were removed from the list before data analysis after checking the genotype for compliance with Hardy Weinberg Equilibrium (HWE) and in order to avoid linkage disequilibrium using PLINK (Purcell, Neale et al. 2007). 13 SNPs were removed for not complying with HWE and an additional 26 SNPs were removed for being in linkage. The entire list of the SNPs removed from the final analyses is colour-coded in the Supplementary table 1. We therefore analyzed the data using a reduced list of 97 SNPs. The statistical significance was set at $p \leq 0.05$ (Balding 2006).

6.1.2.3. SNP Genotype analysis using PLINK software

Analysis of the data for SNP association with PCa based on aggressiveness and gene x environment interaction for risk of PCa were both carried out using PLINK- a tool set for whole genome association (Purcell, Neale et al. 2007). PLINK's clustering approach is based on the genome-wide average proportion of alleles shared IBS between two individuals SNPs,
i.e., pairing up the SNPs based on genetic identity (Purcell, Neale et al. 2007). The IBS clustering is used to test whether the SNPs of two individuals belong to the same population. Following the stratification analysis, we performed a standard case-control association test using a Cochran-Mantel-Haenszel statistic (1 df (degree of freedom)) that tests for SNP-disease association conditional on the clustering. The statistical significance was set at p≤0.05 (Balding 2006).

6.1.2.4. Data Analysis– Stage I: First MLP Process

In this stage, for input to the artificial neural network technique using MLP available in WEKA, an ARFF (Attribute-Relation File Format) file is created and then the first process of MLP is carried out. In this study, a feed forward artificial neural network accompanied by back propagation learning model is applied together with input, output and hidden layers. Two different types of ANN-MLP analysis is conducted in this stage. One analysis is between total PCa (aggressive and non-aggressive) and control patient samples and the other analysis is between aggressive prostate cancer and control patient samples.

6.1.2.4.a. WEKA

WEKA short for Waikato Environment for Knowledge Analysis was formed by the University of Waikato in Hamilton, New Zealand (Hall, Frank et al. 2009). The WEKA platform is a compilation of data pre-processing tools and machine learning methods. It supplies substantial support for the complete course of exploratory data mining, involving developing the input data, assessing learning strategies statistically, and visualizing the input data and the outcome of learning (Hall, Frank et al. 2009; Witten, Frank et al. 2016). WEKA contains a neural network classifier known as MLP that is trained by adopting the concept of back propagation. WEKA (Version 3.8.1) was used to carry out the neural network process by us.
6.1.2.4.b. ANN

ANNs are also known as connectionist models as the association weights signify the memory of the neural structure (Agatonovic-Kustrin and Beresford 2000). The ANN classifier (MLP) within WEKA employs three-layer feed forward neural network accompanied by sigmoid function. This sigmoid function is employed to encode every net as an eight-bit binary unit. ANNs has already been employed for diagnosing and predicting the potential cancer biomarkers (Stephan, Cammann et al. 2002). A feedforward ANN with back propagation learning model is implemented with input, hidden and output layer. The ANN classifier uses a three-layer feed forward neural network with sigmoid function is used to encode each net as an eight bit binary value. The ANN uses genetic algorithm (GA) to reduce the mean squared error of the training set data in the network. The fitness value is selected for maximum parameter to be choosing based on mean square error. The receiver operating characteristic curve (ROC) is used to measure the ratio of false positive and false negative values. The training algorithms used are the LMAM (Levenberg-Marquardt with Adaptive Momentum) and OLMAM (Optimized Levenberg-Marquardt with Adaptive Momentum). The LM method is used for testing the non-linear squares. A feed forward neural network is used where the data for SNP genotyping is divided into four regions basal, parabasal, intermediate and superficial layer (Devi, Ravi et al. 2016).

6.1.2.4.c. Multi Layer Perceptron

Multi Layer Perceptron (MLP) is one of the most frequently employed ANNs. The MLP is a non-linear neural network and is comprised of neurons that are classified in layers. Generally, MLP is comprised of at least three layers, that is, an input layer, an output layer, and one or more hidden layers (Delashmit and Manry 2005). In this study, the MLP neural network is employed to extract weights with higher accuracies. The option of the quantity of hidden layers is a crucial option to be contemplated when developing MLP-ANNs. It was
demonstrated in 1989 by Cybenko that a neural network with a single hidden layer can estimate any continual functions (Cybenko 1989). Nonetheless, a neural network with multiple hidden layers targeting bigger problems was indicated by Popescu, et al., 2009 (Popescu, Balas et al. 2009). This can contribute to the training of the neural network determining for not many local minimums and depletion of the network errors. For these reasons, a dual hidden layer perceptron network is chosen in this study.

The choice of an appropriate activation function for the neurons in the different layers of MLP is very crucial to the performance of the network. The linear activation function is generally used for the input neurons because it transmits the input dataset directly to the next layer with no transformation. The choice of activation function for the output layer neurons is a function of the problem being solved. In our case, which is a multiclass learning problem, we decided to select the sigmoid node function for the output layer neurons because it has the capability to handle either continuous values or $n$-class binary classification problems. The sigmoid node function is also chosen for the neurons in hidden layers because it is nonlinear and differentiable. Differentiability and nonlinearity are vital requirements for MLP training algorithms (Popescu, Balas et al. 2009).

MLP is a feedforward ANN model that maps sets of input data onto a set of suitable outputs and its most significant application has been in the emerging area of artificial intelligence. An MLP constitutes of several layers of nodes in a directed graph. MLP uses a supervised learning approach called backpropagation, as stated earlier, for training the network (Lo Brano, Ciulla et al. 2014).

The output of each node is called its "activation". Weight values are associated with each vector and node in the network, and these values constrain how input data are related to output data. Weight values associated with individual nodes are also known as biases. Weight values are determined by the iterative flow of training data through the network (i.e., weight
values are established during a training phase in which the network learns how to identify particular classes by their typical input data characteristics.

6.1.2.4.d. **ARFF File Creation**

For input to the ANN technique employing MLP, the data for the 623 patients (PCa versus Control) and 566 patients (Aggressive versus Control) was converted into two different ARFF files. In this step, for the first analysis, a first ARFF file was created which contained the SNP data for the 97 SNPs belonging to the 623 patient samples. The SNP data typically coded in biallelic codes such as A/T, C/C, G/C, etc. was converted into 16-bit binary codes, that is, using only the two numeric forms (0 and 1).

Overall, the ARFF file consisted of 623 instances, 1552 attributes, two datatypes (0 and 1) and two classes (PCa and Control). Since there were total of 254 PCa patients recruited (197 aggressive samples and 57 non-aggressive samples) and 369 healthy control samples in this study for the first analysis, the number of instances is 623. And, for the second analysis, a second ARFF file was created which contained the SNP data for the 97 SNPs belonging to 197 aggressive PCa patients and 369 healthy controls. Similar conversion approach of converting biallelic codes into 16-bit binary codes as the first analysis was applied in this step. Overall, the ARFF file consisted of 566 instances, 1552 attributes, two datatypes (0 and 1) and two classes (PCa and Healthy Controls).

6.1.2.4.e. **First Process of MLP**

The first process of MLP (Stage-I) was applied on the two different datasets (i.e. on the 623 and 566 patient samples, respectively) for the 97 SNPs using the training set test option and cross-validation test option with 10 folds. The training set test option is where, 100% of samples are employed for training, whereas, the 10-fold cross-validation test option is where, 90% of samples are employed for training and 10% for cross-validation. The 10-fold cross-
validation test option was employed in this study because its implementation has the potential to resist over-fitting. The following parameters of MLP with two hidden layers; learning rate of 0.3; momentum of 0.2; training time of 150; seeding and validation set size as 0; validation threshold of 20 were chosen in this stage for both the analysis. The results using the training set and 10-fold cross-validation test options of the first process of MLP are shown in Table 6.3.

6.1.2.5. Data Analysis– Stage II: Weights (SNPs) Extraction

In this stage, the weights were first extracted from the first process of MLP and their corresponding SNPs associated with that extracted weights were later extracted based on a set threshold value (more details below). Perceptrons can be trained by a basic learning algorithm that is commonly known as the delta rule. It computes the errors between the computed output data, and employs this to generate an alteration to the weights, therefore applying a framework of gradient descent. Nearly all perceptrons have outputs of 1 or -1 with a threshold of 0 and there is a little evidence that such network can be trained more rapidly than network generated from nodes with distinct activation and deactivation values.

6.1.2.6. Data Analysis– Stage III: Second MLP Process

In this stage, a second ARFF file is created using the extracted weights and then the second process of MLP is carried out on the extracted weights (SNPs). Two different types of ANN-MLP analysis on the similar datasets as Stage I is conducted in this stage. The main purpose of this stage is to see whether the overall cancer/control classification using the training set/10-fold cross validation strategy is still significantly good using the extracted SNPs from both analysis.
6.1.2.6.a. ARFF File Creation

The data for the 623 patients (PCa versus Control) and 566 patients (Aggressive versus Control) was converted into two different ARFF files for input to the second process of MLP. In this step, for the first analysis, a first ARFF file was created which contained the SNP data for the 14 extracted SNPs belonging to the 623 patient samples. As previously stated in section 3.4, the SNP data was converted into 16-bit binary codes. Overall, the ARFF file consisted of 623 instances, 224 attributes, two datatypes (0 and 1) and two classes (PCa and Healthy Controls). Also in this step, for the second analysis, a second ARFF file was created which contained the SNP data for the 14 SNPs belonging to the 566 patient samples. Overall, the ARFF file consisted of 566 instances, 224 attributes, two datatypes (0 and 1) and two classes (PCa and Healthy Controls).

6.1.2.6.b. Second Process of MLP

The second process of MLP (Stage-III) was applied on the two different datasets (i.e. on the 623 and 566 patient samples, respectively) for the 14 extracted SNPs using the training set test option and cross-validation test option with 10 folds. The following parameters of MLP with two hidden layers; learning rate of 0.3; momentum of 0.2; training time of 150; seeding and validation set size as 0; validation threshold of 20 were chosen in this stage for both the analysis. The results of the second process of MLP are shown in Table 6.4. Table 6.4 provides the results for the second process of MLP using the training set and 10-fold cross-validation test options for the first and second analysis.

6.1.3. Results

We analyzed the results for risk association of SNPs to PCa (Table 6.1) and aggressive PCa (Table 6.2) using IBS and ANN. Very interestingly, barring one and three SNPs in the categories of risk for PCa and for aggressive PCa respectively, the results did not tally.
On comparing the data thus generated, we could remove 97% of the SNPs studied by us and brought down the list of high risk biomarkers for aggressive PCa to only three, which is a very manageable number of future studies.

Table 6.1: Results identifying risk of SNPs for PCa using IBS Clustering and ANN

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene Location</th>
<th>Tested Allele</th>
<th>IBS Clustering</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs17793693</td>
<td>PPAR-γ</td>
<td>3p25</td>
<td>A</td>
<td>3.544</td>
<td>1.58454</td>
</tr>
<tr>
<td>rs7931342</td>
<td>MYEOV</td>
<td>11q13</td>
<td>T</td>
<td>1.469</td>
<td>0.00175</td>
</tr>
<tr>
<td>rs5945619</td>
<td>NUDT11</td>
<td>Xp11</td>
<td>C</td>
<td>0.6194</td>
<td>0.00704</td>
</tr>
<tr>
<td>rs10896438</td>
<td>MYEOV</td>
<td>11q13.3</td>
<td>G</td>
<td>0.7208</td>
<td>0.00815</td>
</tr>
<tr>
<td>rs7920517</td>
<td>MSMB</td>
<td>10q11</td>
<td>A</td>
<td>0.7296</td>
<td>0.01141</td>
</tr>
<tr>
<td>rs12529</td>
<td>AKR1C3</td>
<td>10p15</td>
<td>C</td>
<td>0.7485</td>
<td>0.01581</td>
</tr>
<tr>
<td>rs2659056</td>
<td>KLK15</td>
<td>19q13</td>
<td>C</td>
<td>1.341</td>
<td>0.03289</td>
</tr>
<tr>
<td>rs887391</td>
<td>SLC26A6</td>
<td>19q13</td>
<td>C</td>
<td>1.369</td>
<td>0.03516</td>
</tr>
<tr>
<td>rs1799977</td>
<td>MLH1</td>
<td>3p21</td>
<td>G</td>
<td>1.31</td>
<td>0.04179</td>
</tr>
<tr>
<td>rs17632542</td>
<td>KLF3</td>
<td>19q13.33</td>
<td>C</td>
<td>0.5974</td>
<td>0.04538</td>
</tr>
<tr>
<td>rs2242652</td>
<td>TERT</td>
<td>5p15</td>
<td>A</td>
<td>1.367</td>
<td>0.04769</td>
</tr>
<tr>
<td>rs1056827</td>
<td>CYP1B1</td>
<td>2p21</td>
<td>A</td>
<td>-</td>
<td>5.62277</td>
</tr>
<tr>
<td>rs3787268</td>
<td>MMP9</td>
<td>20q12</td>
<td>A</td>
<td>-</td>
<td>3.26000</td>
</tr>
<tr>
<td>rs12621278</td>
<td>ITGA6</td>
<td>2q31</td>
<td>G</td>
<td>-</td>
<td>2.19102</td>
</tr>
<tr>
<td>rs10244329</td>
<td>LEP</td>
<td>7q33</td>
<td>T</td>
<td>-</td>
<td>1.86635</td>
</tr>
<tr>
<td>rs1127678</td>
<td>FASN</td>
<td>17q25</td>
<td>T</td>
<td>-</td>
<td>1.74696</td>
</tr>
<tr>
<td>rs8102476</td>
<td>FASN</td>
<td>19q13</td>
<td>A</td>
<td>-</td>
<td>1.50132</td>
</tr>
<tr>
<td>rs12155172</td>
<td>FASN</td>
<td>7p15</td>
<td>A</td>
<td>-</td>
<td>1.49415</td>
</tr>
<tr>
<td>rs4965373</td>
<td>SEPS1</td>
<td>15q26</td>
<td>A</td>
<td>-</td>
<td>1.40628</td>
</tr>
<tr>
<td>rs10486567</td>
<td>JAZF1</td>
<td>7p15</td>
<td>A</td>
<td>-</td>
<td>1.29133</td>
</tr>
<tr>
<td>rs6823</td>
<td>VDR</td>
<td>12q12</td>
<td>G</td>
<td>-</td>
<td>1.27024</td>
</tr>
<tr>
<td>rs1050450</td>
<td>GPX1</td>
<td>3p21</td>
<td>A</td>
<td>-</td>
<td>1.11820</td>
</tr>
<tr>
<td>rs2121875</td>
<td>FGF10</td>
<td>5p13</td>
<td>A</td>
<td>-</td>
<td>1.05240</td>
</tr>
<tr>
<td>rs632148</td>
<td>SRD5A2</td>
<td>2p23</td>
<td>C</td>
<td>-</td>
<td>1.01708</td>
</tr>
</tbody>
</table>

Common SNPs with significant association with risk of PCa and aggressive PCa with IBS and ANN
Table 6.2: Results identifying risk of SNPs for aggressive PCa using IBS Clustering and ANN.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Gene name</th>
<th>Gene Location</th>
<th>Tested Allele</th>
<th>Odds Ratio</th>
<th>p-Value</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs17793693</td>
<td>PPAR-γ</td>
<td>3p25</td>
<td>A</td>
<td>4.534</td>
<td>0.00017</td>
<td>1.78813</td>
</tr>
<tr>
<td>rs7931342</td>
<td>MYEOV</td>
<td>11q13</td>
<td>T</td>
<td>1.565</td>
<td>0.00074</td>
<td>-</td>
</tr>
<tr>
<td>rs10896438</td>
<td>MYEOV</td>
<td>11q13.3</td>
<td>G</td>
<td>1.4985</td>
<td>0.00232</td>
<td>1.03736</td>
</tr>
<tr>
<td>rs887391</td>
<td>SLC26A6</td>
<td>19q13</td>
<td>C</td>
<td>1.594</td>
<td>0.00509</td>
<td>-</td>
</tr>
<tr>
<td>rs5945619</td>
<td>NUDT11</td>
<td>Xp11</td>
<td>C</td>
<td>1.694</td>
<td>0.00575</td>
<td>-</td>
</tr>
<tr>
<td>rs17632542</td>
<td>KLK3</td>
<td>19q13.33</td>
<td>C</td>
<td>1.998</td>
<td>0.00827</td>
<td>-</td>
</tr>
<tr>
<td>rs7920517</td>
<td>MSMB</td>
<td>10q11</td>
<td>A</td>
<td>1.4</td>
<td>0.01227</td>
<td>-</td>
</tr>
<tr>
<td>rs11228565</td>
<td>MYEOV</td>
<td>11q13</td>
<td>A</td>
<td>1.433</td>
<td>0.02189</td>
<td>-</td>
</tr>
<tr>
<td>rs10244329</td>
<td>LEP</td>
<td>7q33</td>
<td>T</td>
<td>1.557</td>
<td>0.02344</td>
<td>2.10186</td>
</tr>
<tr>
<td>rs6983561</td>
<td>CCHCR1</td>
<td>6p21.3</td>
<td>G</td>
<td>1.383</td>
<td>0.03656</td>
<td>-</td>
</tr>
<tr>
<td>rs2727270</td>
<td>FADS2</td>
<td>11q13</td>
<td>T</td>
<td>1.525</td>
<td>0.04184</td>
<td>-</td>
</tr>
<tr>
<td>rs12529</td>
<td>AKR1C3</td>
<td>10p15</td>
<td>C</td>
<td>1.294</td>
<td>0.04685</td>
<td>-</td>
</tr>
<tr>
<td>rs2659122</td>
<td>KLK3</td>
<td>19q13.33</td>
<td>C</td>
<td>1.345</td>
<td>0.04748</td>
<td>-</td>
</tr>
<tr>
<td>rs1056827</td>
<td>CYP1B1</td>
<td>2p21</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>5.14236</td>
</tr>
<tr>
<td>rs3787268</td>
<td>MMP9</td>
<td>20q12</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>2.42897</td>
</tr>
<tr>
<td>rs12621278</td>
<td>ITGA6</td>
<td>2q31</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>2.31492</td>
</tr>
<tr>
<td>rs1127678</td>
<td>FASN</td>
<td>17q25</td>
<td>T</td>
<td>-</td>
<td>-</td>
<td>2.08939</td>
</tr>
<tr>
<td>rs632148</td>
<td>SRD5A2</td>
<td>2p23</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>1.85507</td>
</tr>
<tr>
<td>rs6823</td>
<td>VDR</td>
<td>12q12</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>1.35304</td>
</tr>
<tr>
<td>rs10486567</td>
<td>JAZF1</td>
<td>7p15</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>1.33551</td>
</tr>
<tr>
<td>rs1050450</td>
<td>GPX1</td>
<td>3p21</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>1.30895</td>
</tr>
<tr>
<td>rs13254738</td>
<td>PRNCR1</td>
<td>8q24</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>1.19199</td>
</tr>
<tr>
<td>rs12155172</td>
<td>JAZF1</td>
<td>7p15</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>1.02409</td>
</tr>
<tr>
<td>rs2292884</td>
<td>MLPH</td>
<td>2q37</td>
<td>G</td>
<td>-</td>
<td>-</td>
<td>1.02197</td>
</tr>
</tbody>
</table>

Common SNPs with significant association with risk of PCa and aggressive PCa with IBS and ANN

In total, the first process of multilayer perceptron (MLP) for the first analysis generated 3104 weights. 1552 each for cancer and control samples, respectively. Maximum weights for the cancer and control samples were 5.62 and 4.59, respectively. Minimum weights for the cancer and control samples were -3.70 and -3.17, respectively. The average weights for the cancer and control samples were 0.0061 and 0.0019, respectively, whereas, the standard deviation values for the cancer and control samples were 0.383 and 0.309, respectively. A threshold value based on the standard deviation evaluations for the potential weights.
extraction was set to over one. Overall, based on the set threshold value of over one, for first analysis, 19 SNPs were extracted from the cancer weights and 14 SNPs were extracted from the control weights. All the duplicate SNPs were eliminated from this process. In total, 19 SNPs were removed and 14 SNPs were retained in Stage II. Table 6.2 gives the list of the 14 SNPs extracted in this stage for the first analysis. Also, in total, the first process of MLP for the second analysis generated 3104 weights as the total number of SNPs remained the same in both the analysis. Maximum weights for the cancer and control samples were 5.14 and 5.12, respectively. Minimum weights for the cancer and control samples were -2.67 and -2.86, respectively. The average weights for the cancer and control samples were 0.0027 and 0.0044, respectively, whereas, the standard deviation values for the cancer and control samples were 0.362 and 0.362, respectively. Overall, based on the set threshold value of over one, for second analysis, 20 SNPs were extracted from the cancer weights and 16 SNPs were extracted from the control weights. All the duplicate SNPs were eliminated from this process. In total, 22 SNPs were removed and 14 SNPs were retained in this stage. Table 6.1 also gives the list of the 14 SNPs extracted in this stage for the second analysis.

Analysis for risk of PCa using training set option, MLP correctly classified 623 samples for 97 SNPs with an overall accuracy of 99.8395%. PCa samples were correctly classified with an accuracy of 100% and control samples were correctly classified with an accuracy of 99.73%. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 623 samples for 97 SNPs were all 99.8%, respectively. Analysis for aggressive PCa using training set option MLP correctly classified 566 samples for 97 SNPs with an overall accuracy of 99.47%. The PCa samples were correctly classified with an accuracy of 98.48% and control samples were correctly classified with an accuracy of 100%, respectively. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 566 samples for 97 SNPs were 99.5% each (Table 6.3).
Analysing for risk of PCa using a 10-fold cross validation (i.e. 90% training and 10% testing strategy) option MLP correctly classified 623 samples for 97 SNPs with an overall accuracy of 87.9615%. The PCa samples were correctly classified with an accuracy of 83.46% and control samples were correctly classified with an accuracy of 91.06%. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 623 samples for 97 SNPs were all 88%. Analysis for risk of aggressive PCa using a 10-fold cross validation option MLP correctly classified 566 samples for 97 SNPs with an overall accuracy of 86.5724%. The aggressive PCa samples were correctly classified with an accuracy of 76.65% and control samples were correctly classified with an accuracy of 91.87%. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 566 samples for 97 SNPs were 86% each (Table 6.3).
Table 6.3: Results of the first process of MLP (Stage-I) for the first and second analysis

<table>
<thead>
<tr>
<th>Class</th>
<th>ROC Area</th>
<th>P-value</th>
<th>TPR</th>
<th>TNR</th>
<th>FPR</th>
<th>1-ppv</th>
<th>1-tnr</th>
<th>1-fpr</th>
<th>TPR_50</th>
<th>TNR_50</th>
<th>FPR_50</th>
<th>1-ppv_50</th>
<th>1-tnr_50</th>
<th>1-fpr_50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
</tbody>
</table>

Second Analysis - Aggressive PCA Versus Control Samples

<table>
<thead>
<tr>
<th>Class</th>
<th>ROC Area</th>
<th>P-value</th>
<th>TPR</th>
<th>TNR</th>
<th>FPR</th>
<th>1-ppv</th>
<th>1-tnr</th>
<th>1-fpr</th>
<th>TPR_50</th>
<th>TNR_50</th>
<th>FPR_50</th>
<th>1-ppv_50</th>
<th>1-tnr_50</th>
<th>1-fpr_50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
<tr>
<td>Control</td>
<td>0.969</td>
<td>0.000</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
<td>0.969</td>
<td>0.969</td>
<td>0.031</td>
</tr>
</tbody>
</table>
Table 6.4 shows that for the first analysis using training set option MLP correctly classified 623 samples for 97 SNPs with an overall accuracy of 95.0241%. Cancer samples were correctly classified with an accuracy of 93.7% and control samples were correctly classified with an accuracy of 95.94%, respectively. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 623 samples for 97 SNPs were each 95%. And for the second analysis using training set option MLP correctly classified 566 samples for 97 SNPs with an overall accuracy of 91.6961%. Cancer samples were correctly classified with an accuracy of 84.26% and control samples were correctly classified with an accuracy of 95.66%, respectively. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 566 samples for 97 SNPs were each 92%.

Table 6.4 also shows that for the first analysis using 10-fold cross validation (i.e. 90% training and 10% testing strategy) option MLP correctly classified 623 samples for 97 SNPs with an overall accuracy of 85.2327%. PCa samples were correctly classified with an accuracy of 85.04% and control samples were correctly classified with an accuracy of 85.37%, respectively. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 623 samples for 97 SNPs were all 85%, respectively. And for the second analysis using 10-fold cross validation option MLP correctly classified 566 samples for 97 SNPs with an overall accuracy of 84.4523%. Cancer samples were correctly classified with an accuracy of 73.1% and control samples were correctly classified with an accuracy of 90.52%, respectively. The overall true positive rate, precision, recall, and F-measure (all in percentages) of the 566 samples for 97 SNPs were all 84%, respectively.

The results from Table 6.4 signify that the overall cancer/control classification using the training set per 10-fold cross-validation strategy is still significantly good using the extracted SNPs from both analysis.
Table 6.4 Results of the second process of MLP (Stage-III) for the first and second analysis

182


6.1.4. Discussion

One of the major requirements for any case-control study is strict data analysis and interpretation. The data and the analysis methods discussed herewith bring forth several interesting attributes.

One of the most prominent finds of this comparative analysis of SNP data in the backdrop of a case-control study is the difference in the list of SNPs associated with PCa and aggressive PCa that are identified between PLINK and ANN (Figure 6.3 illustrates the findings in aggressive PCa). ANN is a much advanced deep-learning tool, and interestingly, the SNPs that were identified as being statistically significant with varying significant p-values were defined as significant based on weights by ANN, but these SNPs are identified towards the lower half of the table (based on descending weight distribution). We believe this is a very good proof that while dealing with such vast and distributed data, machine-learning along with the established approach is a crucial aspect of extracting the information that could otherwise be ignored and/or missed.

The fact that SNPs rs10244329 and rs17793693 in two genes- LEP and PPAR-γ respectively pertaining to fat metabolism and risk diabetes mellitus (Labayen, Ruiz et al. 2011; Derosa and Maffioli 2012) and SNP rs10896438 in MYEOV- a putative oncogene were shortlisted from a considerable number of SNPs studied as risk for aggressive PCa in our cohort is itself a major victory in narrowing down the number of SNPs associated with this life-threatening disease to a manageable number. Interestingly, Chung et al., (2012) have reported, using in silico genomic analysis that the SNP rs10896438 is present in a recombination hotspot, which localizes to an alternative gene Two pore segment channel 2 (TPCN2) (Chung, Boland et al. 2012).
Previously we have reported a risk association of high body mass index (BMI) and high rates of tobacco smoking (Vaidyanathan, Naidu et al. 2017), especially with aggressive PCa (compared to healthy controls).

Figure 6.3: SNPs significantly associated with risk of aggressive PCa using IBS and ANN analyses.

Although the role of tobacco smoking and/or cessation of tobacco smoking has not yet been established with regards obesity and changes in body weight, these findings of high BMI, high tobacco smoking, and subsequent association of genes pertaining to obesity are in line with the various studies carried out with regards the controversial role of tobacco smoking and obesity with aggressive PCa (Klesges, Meyers et al. 1989; Potter, Pederson et al. 2004; Zbikowski, Jack et al. 2011; Dare, Mackay et al. 2015).

Also of interest is the fact that diabetes mellitus has previously been associated with PCa, such that patients with diabetes mellitus have been suggested to be at a lower risk of PCa (Giovannucci, Harlan et al. 2010). It was previously suggested by us that BMI having a
strong association not only with diabetes mellitus but also with insulin resistance (Al-Goblan, Al-Alfi et al. 2014) and with PCa (Cao and Ma 2011), and hence obesity (and high BMI) is linked with the risk of diabetes mellitus and PCa alike, and it is not diabetes mellitus that has a direct relation with PCa (Vaidyanathan, Naidu et al. 2017). The findings from this exercise are in line with this theory. This will help establish various large-scale analyses and define the role of obesity and/or high BMI on risk of advanced chronic diseases such as aggressive PCa.

The other interesting find is the fact that certain SNPs are identified to be common between the risk of PCa and aggressive PCa using either models, but they do not tally completely indicating that the SNPs in certain genes can play a much more specific role with regards the prognosis of the disease, and certain SNPs may be restricted to only the incidence of the disease.
6.2. Chapter summary

Artificial Neural Networks is undisputedly a “next generation technology” in using a “Computer brain” to define the hidden laws of various challenging associations and/or interactions of SNPs and risk of diseases such as PCa. In our previous section (chapter 5-SNP genotyping), we identified certain results which partly surprised us, the findings of this chapter, however, not only surprised us, but amazed us as well. The underlying story of fat metabolism and risk of PCa and its aggressive form as well as the role of a putative oncogene was decoded by employing this cutting-edge technology.

Having identified a fewer number of genes (with SNPs) that were of much interest to our story, we went ahead and validated our results by a protein expression analysis (Section 7.1) in imported PCa cell-lines to see if this trend can be seen in samples not local to New Zealand.
CHAPTER SEVEN

Protein Expression Analysis

"Every know fact in natural science was divined by the presentiment of somebody, before it was actually verified."

- Ralph Waldo Emerson
Chapter 7. Protein Expression Analysis

Chapter preamble

With results pointing towards a definite gene x environment interaction and associated risk of aggressive PCa among the cohort we studied (New Zealand Caucasian men), we wanted to validate if the results can be defined in imported cell-lines as well, as the results will have much power if replicated in any population worldwide.

We wanted to check for the expression levels of the proteins in different PCa cell-lines such that it was a “blind-study” (Section 7.1). From literature search, we knew that PC-3 is an aggressive PCa cell line, and we wanted to test the protein expression levels of MYEOV, and PPAR-γ to see if they corresponded with aggressiveness of the disease.
7.1. Role of MYEOV-PPARγ-ALK nexus and tobacco smoking and obesity in the risk of prostate cancer- a New Zealand model

Abstract: The gene x environment impact on incidence and progression of diseases is very unique and fascinating, but at the same time generates hope for the many patients suffering with chronic diseases such as prostate cancer (PCa). We have performed a multi-directional study herewith and have identified a very unique and not yet reported link between three genes - MYEOV, ALK, PPAR-γ and the risk of PCa using a case-control study and validation using protein expression in imported cell lines. The result of this unique demographic and molecular analysis on risk of PCa and the potential role of MYEOV, ALK, PPAR-γ is herewith detailed.

7.1.1. Introduction

Prostate cancer (PCa) is the second-most commonly diagnosed cancer among men worldwide (Torre, Bray et al. 2015). In the year 2012, approximately 1.1 million men were diagnosed with PCa, which accounted for 15% of all the new cancer cases registered in men worldwide (Ferlay, Soerjomataram et al. 2015). The incidence rates of PCa vary significantly across the globe with the highest rate recorded in the Australia/ New Zealand (NZ) region, followed by North America, Western and Northern Europe, and some Caribbean Countries. The lowest rates are recorded in Asian countries (Torre, Bray et al. 2015).

PCa is the most commonly registered male cancer and the third-most common cause of cancer deaths in NZ men. It accounted for 27.2% of all male cancer registrations and 13.4% of male cancer deaths in 2013. In the year 2013, 3129 men in NZ were diagnosed with PCa, corresponding to a rate of 953 per million, and 647 deaths were due to PCa with an incidence rate of 174 per million (Ministry of Health 2016).
There have been a number of research-based approaches reported with regards to identification of a biomarker for PCa, including genome-wide association studies (GWAS) (Amin Al Olama, Kote-Jarai et al. 2013) and -omics studies (Prensner, Rubin et al. 2012; Crawford, Ventii et al. 2014) and more are being proposed and looked into (Sharma, Zargar-Shoshtari et al. 2015; McGrath, Christidis et al. 2016). Single nucleotide polymorphisms (SNPs) are the most commonly observed variations in the genome and are often employed for case-control studies to define the role of genetics in human diseases, including PCa (Vaidyanathan, Naidu et al. 2017).

We have reported a number of SNPs as significantly associated with the risk of PCa in a New Zealand cohort (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2012; Karunasinghe, Lange et al. 2013; Bishop, Han et al. 2016; Karunasinghe, Zhu et al. 2016), and even aggressive PCa (Vaidyanathan, Naidu et al. 2017), in the recent past. However, a holistic approach had not been considered yet with the employment of statistical approaches to define the role of environmental factors for the risk of PCa. Certain external factors such as high body mass index (BMI), alcohol consumption, and smoking tobacco were recently collectively analyzed and discussed as risk for PCa by us (Vaidyanathan, Naidu et al. 2017), the specific gene x environment functions, however, is yet to be defined clearly with regards certain lifestyle factors and risk of the disease.

We identified a significant association with tobacco smoking, and high BMI and with the SNPs present in a putative oncogene- *MYEOV* (*Myeloma Overexpressed*) and the risk of PCa in our cohort (Vaidyanathan, Naidu et al. 2017). This was an interesting find as it is often suggested that *MYEOV* expression can be inhibited by deoxyribonucleic acid (DNA) -methylation (Janssen, Imoto et al. 2002), and tobacco smoking causes differential DNA methylation (Lee and Pausova 2013), and tobacco smoking can also have an effect on overall gain or loss of body weight. These parameters can in turn affect the expression of certain
genes (Meng, Zhu et al. 2017). We have herewith used the data already reported by SNP genotyping in our New Zealand cohort and have explored the opportunities to identify the effect of tobacco smoking on MYEOV and ALK (Anaplastic Lymphoma Kinase) and obesity on PPAR-γ (Peroxisome Proliferator-Activated Receptor gamma) and the epigenomic factors which may affect the expression of the genes of interest thereby causing aggressiveness of the disease. Previous studies from our group have identified a risk association of tobacco smoking with PCa in our cohort (Karunasinghe, Han et al. 2013). We believe that tobacco smoking may have a wider implication than just risk for lung cancers (IARC 2004). Bearing this in mind, we investigated into the close proximity of the main oncogene of our interest—MYEOV, and certain other genes significantly associated with risk of PCa in our cohort and including the gene ALK. The gene ALK is involved with non-small lung carcinoma (Gadgeel, Shaw et al. 2017). Beyond, we also carried out protein expression studies to prove the expression of these important proteins in imported cell lines. Thus, a multi-disciplinary approach was needed to be carried out to investigate further. This is a first study of this kind to propose the role of MYEOV, ALK and PPAR-γ and the implication of tobacco smoking and high BMI in the occurrence of PCa in a New Zealand population.

7.1.2. Materials and Methods

7.1.2.1. Statistical analysis of demographics factor and SNP genotype data (association and interaction):

The variation of lifestyle characteristic- tobacco smoking between pathology was tested using the Fisher exact test (Vaidyanathan, Naidu et al. 2017) and high BMI using Mann-Whitney U test.

SNP genotyping was carried out on 136 SNPs, based on literature search of the published genome-wide association studies (GWAS) for both PCa and its aggressive form, as mentioned in Vaidyanathan et al., 2017 (Vaidyanathan, Naidu et al. 2017) in 254 patients.
with PCa and 369 healthy controls either by SEQUENOM MassARRAY iPLEX® assay (2006) or by TaqMan® SNP genotyping assay using the manufacture’s protocols (Vaidyanathan, Naidu et al. 2017). We have used the data already reported by us to analyse the risk of PCa, without any classifications into its aggressive and/or non-aggressive forms (Thompson, Thrasher et al. 2007), as defined by D’Amico et al., (1998) (D'Amico, Whittington et al. 1998), as we wanted to see the effect of gene x environment interactions (between tobacco smoking and risk of PCa) without any bias and loss of data due to the classifications.

The data that was generated using SNP genotyping was analyzed using PLINK software (Purcell, Neale et al. 2007). SNP genotyping was done for a total of 136 SNPs, but after checking for compliance with Hardy Weinberg Equilibrium (HWE), and in linkage disequilibrium, 97 SNPs were employed for the final analysis (Vaidyanathan, Naidu et al. 2017). The HWE and linkage analyses were done by employing PLINK software version 1.07 (Purcell, Neale et al. 2007).

A structured association approach providing a simple but powerful method to detect population stratification is implemented in the PLINK-based analysis (Purcell, Neale et al. 2007). PLINK’s clustering approach is based on the genome-wide average proportion of alleles shared identical-by-state (IBS) between two individuals, i.e., pairing up the SNPs based on genetic identity rather than descent (Purcell, Neale et al. 2007). The IBS clustering is used to test whether the SNPs of two individuals belong to the same population. Following the stratification analysis, we performed a standard case-control association test using a Cochran-Mantel-Haenszel statistic (1 df (degree of freedom)) that tests for SNP-disease association conditional on the clustering (Purcell, Neale et al. 2007). We used \( p \leq 0.05 \) to denote statistical significance. Correction for multiple testing was applied to the analyzed data obtained, so as to maintain the linearity of genotype-phenotype relationship (Balding
2006). As the tested SNPs are already proven as associated with PCa incidence by other researchers (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2012; Karunasinghe, Lange et al. 2013; Bishop, Han et al. 2016; Karunasinghe, Zhu et al. 2016; Vaidyanathan, Naidu et al. 2017), variations that showed significance before Bonferroni correction were also considered for discussion in the current study.

The covariates included the environmental factor - tobacco smoking history. Also three genotypic models: additive (ADD), dominant deviation (DOMDEV), and general_2 df joint test of both additive and dominant deviation (GENO_2DF) were generated. An additive model represents the additive effects of SNPs i.e., the effect of each additional minor allele as represented by the direction of the regression coefficient. For example, a positive regression coefficient indicates that the minor allele increases risk. A DOMDEV model represents a separate test of the dominance component, and a general model represents the joint test of both ADD and DOMDEV components. However, in contrast to a dominance model, ADD refers to a variable coded in such a way (0, 1, 0 for three genotypes AA, Aa, aa) that it represents the dominance deviation from additivity without specifying whether a particular allele is dominant or recessive. Effects of genotype x environmental (tobacco smoking) interactions was tested and eventually adjusted for.

7.1.2.2. Phylogenetic tree construction:

Since the genes of interest with regards to our project have been identified to be significantly associated with the risk of aggressive prostate cancer, we believed that there was enough evidence to look into the proximities of the same for further research. We therefore plotted a phylogenetic tree to achieve the same. Sequence alignment method (Needleman and Wunsch 1970) was used to construct the cladistics-based phylogenetic tree to decipher the existence of significant homology between the genes of interest identified by using PLINK analysis in our case-control study. The four prominent steps that were used for this are
Assemblaging of the sequence data in FASTA format (Lipman and Pearson 1985), followed by alignment and determination of the substitution model, and Tree construction and finally Tree estimation. In addition to genes considered with SNP data, we also considered the ALK gene in this construction due to the negative influence on risk of certain cancers, and the high incidence of tobacco smokers and the unconfirmed impact of tobacco smoking and obesity expecting no co-relation to prove the relation between risk of tobacco smoking and PCa (with ALK as a negative control of sorts).

**7.1.2.3. Western blot analyses:**

**7.1.2.3.a. Cell lines:**

The three PCa cell lines—PC-3 (ATCC® CRL-1435™), DU 145 (ATCC® HTB-81™) and LNCaP (ATCC® CRL-1740™), and two ovarian cancer cell lines SK-OV-3 (ATCC® HTB-77™) and OVCAR3, and skin cancer cell line A345, used for this study, were obtained from American Type Culture Collection (ATCC) and cultured in F-12K (Thermofisher, USA); Eagle’s Minimum Essential Medium (Thermofisher, USA); and RPMI 1640 (Thermofisher, USA) respectively, plus 10% fetal bovine serum (FBS) (Thermofisher, USA) was added and cultured in a humidified incubator with a 5% CO₂ atmosphere at 37°C.

**7.1.2.3.b. Western blot assay:**

The cells from the lines PC-3, DU 145, LNCaP, SK-OV-3, OVCAR3 and A345 were harvested and lysed with buffer supplemented with protease inhibitors. Equal amounts of protein (40 μg protein each lane) were separated by 12% SDS-PAGE and transferred to polyvinylidene difluoride membranes (Millipore, USA). After incubation in blocking buffer TBS (0.025 mM tris-base, 0.1 M NaCl, pH 7.4, 2 % BSA) for 1 hour at room temperature, the membranes were incubated with primary antibodies against ALK, MYEOV, PPAR-γ at 4°C overnight. After washing with TBST (0.05% tween 20, 0.025 mM tris-base, 0.1 M NaCl,
pH 7.4) for 5 times, with 6 minutes intervals, the membranes were incubated with horseradish peroxidase-conjugated secondary antibodies at room temperature for 1 hour.

After the incubation, the membranes were washed again with TBST for 5 times with 6 minutes interval. The membranes were then incubated with SuperSignal® West Pico Chemiluminescent Substrate (Thermofisher, USA) for 4 minutes. Protein bands in the membrane were then taken onto an x-ray film, which was then developed in a VersaDoc (BioRad) imaging system.

7.1.2.3.c. Quantification of protein:

The quantification of protein band density was carried out using a Bio-Rad GS-710 Calibrated Imaging Densitometer. Films containing protein bands were scanned and quantified using Quantity One software (BioRad). Relative density of the bands was normalized to β-actin as the loading control.

7.1.3. Results and Discussion

7.1.3.1. BMI and pathology:

Since there is a large number of ever tobacco smokers and overweight population in New Zealand (2015), we carried out a Fisher exact test and Mann-Whitney U test respectively to analyse the risk of PCa due to a positive correlation with these lifestyle factors among the patients recruited and healthy controls for our study. The BMI status is provided in the Table 7.1 and the tobacco smoking (ever smokers) status is provided in the

Table 7.2.
We found strong association between high BMI (Table 7.1) and tobacco smoking (Table 7.2) and the risk of PCa in our cohort, and hence in our subsequent analysis of SNP genotyping data, for gene x environment interactions, we statistically adjusted for these parameters.

### 7.1.3.2. Genetic polymorphism variations and risk of prostate cancer:

Out of the 97 SNPs studied, 11 SNPs were identified to be significantly associated with risk of PCa in our patient cohort when compared with healthy controls without any statistical adjustments and 7 SNPs after adjusting for tobacco smoking only (Table 7.3).
The Table 7.3 comprises the results of the statistically significant SNPs associated between patients with PCa and the healthy controls assessed before and after adjusting for tobacco smoking. The results that were identified as statistically significant before the use of the conservative Bonferroni correction has been discussed here (Vaidyanathan, Naidu et al. 2017)

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene Location</th>
<th>SNP ID</th>
<th>Tested Allele</th>
<th>Before adjustment</th>
<th>After adjustment for BMI</th>
<th>After adjustment for tobacco smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-Value</td>
<td>OR</td>
<td>p-Value</td>
</tr>
<tr>
<td>PPAR-γ</td>
<td>3p25</td>
<td>rs17793693</td>
<td>A</td>
<td>0.000145</td>
<td>3.544</td>
<td>--</td>
</tr>
<tr>
<td>MYEOV</td>
<td>11q13</td>
<td>rs7931342</td>
<td>T</td>
<td>0.001745</td>
<td>1.469</td>
<td>0.000364</td>
</tr>
<tr>
<td>NUDT11</td>
<td>Xp11</td>
<td>rs5945619</td>
<td>C</td>
<td>0.007039</td>
<td>0.6194</td>
<td>0.000364</td>
</tr>
<tr>
<td>MYEOV</td>
<td>11q13.3</td>
<td>rs10896438</td>
<td>G</td>
<td>0.00815</td>
<td>0.7208</td>
<td>0.000364</td>
</tr>
<tr>
<td>MSMB</td>
<td>10q11</td>
<td>rs7920517</td>
<td>A</td>
<td>0.01141</td>
<td>0.7296</td>
<td>0.003425</td>
</tr>
<tr>
<td>AKR1C3</td>
<td>10p15</td>
<td>rs12529</td>
<td>C</td>
<td>0.01581</td>
<td>0.7485</td>
<td>0.04263</td>
</tr>
<tr>
<td>KLF15</td>
<td>19q13</td>
<td>rs2659056</td>
<td>G</td>
<td>0.03289</td>
<td>1.341</td>
<td>0.04263</td>
</tr>
<tr>
<td>SLC26A6</td>
<td>19q13</td>
<td>rs887391</td>
<td>C</td>
<td>0.03516</td>
<td>1.369</td>
<td>0.02584</td>
</tr>
<tr>
<td>MLH1</td>
<td>3p21</td>
<td>rs1799977</td>
<td>G</td>
<td>0.04179</td>
<td>1.31</td>
<td>0.04179</td>
</tr>
<tr>
<td>KLF3</td>
<td>19q13.3</td>
<td>rs17632542</td>
<td>C</td>
<td>0.04538</td>
<td>0.5974</td>
<td>0.04538</td>
</tr>
<tr>
<td>TERT</td>
<td>5p15</td>
<td>rs2242652</td>
<td>T</td>
<td>0.04769</td>
<td>1.367</td>
<td>0.04769</td>
</tr>
<tr>
<td>MYEOV</td>
<td>11q13</td>
<td>rs11228565</td>
<td>A</td>
<td>0.03484</td>
<td>0.5871</td>
<td>0.02221</td>
</tr>
<tr>
<td>Sep-15</td>
<td>1p22.3</td>
<td>rs5845</td>
<td>T</td>
<td>0.03484</td>
<td>0.5871</td>
<td>0.02221</td>
</tr>
</tbody>
</table>

SNPs with significant association with risk of PCa before and after adjusting for tobacco smoking and BMI

Large scale genetic association studies such as GWAS have come up with direct SNP associations and risk of PCa. However, SNP interactions with demographic and lifestyle factors could also add to the allelic effect producing a modified risk of a disease, especially with the focus of research shifting from genes and genomics to gene x environment interactions (Vaidyanathan, Naidu et al. 2017). As mentioned above, because the role of external factors such as tobacco smoking is well established in causing epigenetic changes
(Breitling 2013; Ambatipudi, Cuenin et al. 2016), and which could in turn lead to the regulation of the putative oncogene, *MYEOV*, which is shown to be regulated by epigenetic factors via a DNA-methylation mechanism (Janssen, Imoto et al. 2002). This was a major shift in our approach for the next step of experiment and data analysis and led us to statistically adjust our data for tobacco smoking.

The SNPs that are identified as statistically significantly associated with risk of PCa when compared with healthy controls in our population after adjusting for tobacco smoking are rs5845 present on the gene *SEP15*; rs12529 present on the gene *AKR1C3* present at the chromosomal location 10p15; rs7920517 present in the gene *MSMB* near 10q11; rs7931342 and rs11228565 found in the gene *MYEOV* at 11q13 region; and rs887391 found in the gene *SLC26A6* near 19q13 region. These SNPs were identified to be statistically significantly associated with risk of aggressive PCa in our cohort in our previous study as well (Vaidyanathan, Naidu et al. 2017). Interestingly, when the parameter of aggressiveness was replaced for risk of PCa in general, we identified two other SNPs- rs2242652 located near the 5p15 region near the gene *TERT* and rs2659056 located near the 19q13 region near the gene *KLK15* which were not previously reported by us as risk for the disease. This knowledge is important, as we could have easily missed to identify the implication of these SNPs as biomarker(s) for PCa, if we did not explore further. The finding is interesting, because gene *KLK3*, also known as *PSA* (present biomarker for PCa), has shared protein domains as the gene *KLK15* and co-expresses with the gene *TERT*.

In context of the present study, it was interesting to find that SNPs in the putative oncogene- *MYEOV* remained significant even after the statistical adjustments for BMI and for tobacco smoking; and also, the genes pertaining to prostate well-being- *MSMB* and *AKR1C3*, along with the SNP in the gene *SLC26A6*. This indicates that these polymorphisms have a strong correlation with the risk for PCa with and without the influence of personal lifestyle.
factors such as tobacco smoking and the influence of high BMI. This is an important finding, as we are the first to report the combination of these SNPs (and, genes) in a case-control study as a risk for PCa in a New Zealand population of European ethnicity. It will be interesting to see the influence of the afore SNPs as it seems to have an impact on the risk of PCa with and without adjusting for environmental factors. This justifies a strong genomic relation with the aggressive disease, but it would seem that environmental factors too may have some influence, may be in exacerbating the problem.

7.1.3.3. Cladistics-based phylogenetic tree:

Since the expression of the putative oncogene \textit{MYEOV} has been defined to be regulated by epigenetic factors (Janssen, Imoto et al. 2002; Papamichos, Margaritis et al. 2015), which in turn could be influenced by lifestyle choices such as tobacco smoking, and tobacco smoking was identified as a risk for PCa in our cohort too, we looked into one of the most studied and targeted genes for therapy for lung cancer, a major disease influenced by tobacco smoking (Spiro and Silvestri 2005; Pesch, Kendzia et al. 2012)– ALK (Shaw and Solomon 2011; Selinger, Rogers et al. 2013; Liao, Lin et al. 2015). The gene \textit{ALK} has been found to be rearranged, mutated, or amplified in non-small cell lung cancer (Calio, Bria et al. 2016), and in PCa (Orozco, Goh et al. 2013) including fusion gene products. We explored the possibility of shared history among the genes identified on the basis of SNP genotyping as risk for PCa with the ALK gene using a neighbor-joining phylogenetic tree (Figure 7.1).

The cladistics-based phylogenetic tree generated by us indicated that the genes \textit{MYEOV} and \textit{ALK} had shared history, and two of its closest neighbors were the genes \textit{PPAR-\gamma} and \textit{MSMB}. These results are interesting as rs17793693 present near the gene \textit{PPAR-\gamma} was identified as one of the most significantly associated risk SNP for PCa by us before statistical adjustments for tobacco smoking (effect of gene x environment interaction), as is represented in Table 7.1. \textit{MYEOV} is established as human-specific \textit{de novo} protein-coding oncogene, and
$ALK$ can also be oncogenic, by one of three ways- forming a fusion gene, or gaining additional gene copies, or mutations of the actual DNA code of the gene in general, which all could lead to uninhibited or perpetual kinase activity.
Figure 7.1 Cladistics-based phylogenetic tree.
There are two parts to the results we identified after plotting the phylogenetic tree and each interesting in its own way. The evidence that MYEOV and ALK are so close, points towards the various theories we have been proposing with regards the risk of the genes being influenced due to tobacco smoking and risk of cancers. Second, the result obtained by plotting the phylogenetic tree indicates as to why there is a correlation between smoking tobacco and obesity (Dare, Mackay et al. 2015) and low and/or poor sperm count (Kovac, Khanna et al. 2015).

One of important genes identified by our current and previous studies as a risk for PCa is PPAR-γ (Vaidyanathan, Naidu et al. 2017) is present mainly in the adipose tissues (Segawa, Yoshimura et al. 2002) and has an established role in obesity (Lv, Zhang et al. 2017) and diabetes (Derosa and Maffioli 2012; Lv, Zhang et al. 2017). Although the role of tobacco smoking and/or cessation of tobacco smoking has not yet been established with regards obesity and changes in body weight, this finding is in line with the various studies carried out with regards the controversial role of tobacco smoking and obesity (Klesges, Meyers et al. 1989; Potter, Pederson et al. 2004; Zbikowski, Jack et al. 2011; Dare, Mackay et al. 2015). The identification of PPAR-γ, a gene pertaining to obesity so closely related to ALK, a gene that is much interest with regards to non-small cell lung cancer will perhaps help researchers working in this field of study to better understand the role of tobacco smoking on obesity.

It is well established that cessation of tobacco smoking improves reproductive outcomes in males (Harlev, Agarwal et al. 2015; Kovac, Khanna et al. 2015). It is suggested that tobacco smoking affects sperm quality (Mostafa 2010), and very interestingly, the protein MSMB, one of the three prominent proteins secreted by human prostate gland (Lilja and Abrahamsson 1988), has been suggested to play a role in the process of fertilization (Anahi Franchi, Avendano et al. 2008) and is associated with the sperm surface (Anahi Franchi, Avendano et al. 2008). One of the other two proteins secreted by the human prostate gland is
prostate-specific antigen (PSA) (Lilja and Abrahamsson 1988), which is measured to identify the possibility of PCa incidence (D'Amico, Whittington et al. 1998). Although the number of patients identified early with PCa increased since the use of PSA as a diagnostic tool (Ilyin, Belkowski et al. 2004), yet it is not a specific marker and led to unnecessary biopsies in a number of cases (Gray, Delahunt et al. 2004); however, factors such as tobacco smoking may alter the levels of PSA (Gelmann, Chia et al. 2001; Escandriolo Nackauzi, Colla et al. 2012; Koc, Akgul et al. 2013), thereby misleading clinical diagnostics.

These findings with regards the genes pertaining to obesity and sperm quality so closely related to MYEOV and ALK- a kinase which plays an important role in various cellular pathways and is crucial in metabolism, cell signaling, and even protein regulation, the phylogenetic tree made us wonder if tobacco smoking was the major factor triggering the response from the genes PPAR-γ and MSMB respectively and thus being identified as significantly associated with risk of PCa. When we refereed back to the results obtained from SNP data, it was no surprise when we found that the statistical significance of the gene PPAR-γ to PCa was lost when we adjusted for tobacco smoking. The gene MSMB remained statistically significant even after adjusting for tobacco smoking.

With regards the nexus between the genes ALK and MYEOV and tobacco smoking and risk of PCa, studies have found that the ALK-lung cancers are observed more commonly among the light cigarette smokers and/or non-smokers, but is well established that a significant number of the patients with this disease are current or former cigarette smokers (Rudin, Avila-Tang et al. 2009). This is in line with our findings for the association between the SNPs in the putative oncogene MYEOV and a risk of PCa in that the gene is found to be statistically significant both, before and after statistical adjustment for tobacco smoking. Moreover, it is well established that people change their lifestyle after being diagnosed for any chronic illness such as PCa, as was reported for alcohol consumption (Vaidyanathan, Naidu et al. 2013).
This also supports our gene of target \textit{ALK} for the cladistics-based phylogenetic tree discussed herewith. We are conscious of the fact that we have produced a cladistics-based phylogenetic tree only for the genes of our interest, and if the tree is generated using all the genes identified by the human genome project, the spread of the genes may not be the same. However, our results still hold tremendous value because, there is no general agreement about the choice of the total number of genes and a phylogenetic tree with significant results (Chen, Deng et al. 2016).

\subsection{Protein expression levels:}

With interesting results with regards the correlation between the genes \textit{PPAR-\textgamma}, and \textit{MYEOV}, in terms of association as a risk factor for PCa in our analyses (Table 7.1), and the shared evolutionary history among the genes \textit{MYEOV} and \textit{ALK}, as was identified by plotting a cladistics-based phylogenetic tree (Figure 7.1), we decided to define the protein expression levels of \textit{ALK}, \textit{MYEOV} and \textit{PPAR-\textgamma} in three PCa cell lines- PC-3, DU145, LNCaP, and three cancer cell lines (as a comparative cancer control)- SK-OV-3 and OVCAR3 (ovarian cancer cell lines), and A431 (skin cancer cell line) (Figure 7.2). PC3 is the more aggressive line amongst the afore mentioned three PCa cell lines with a morphology similar to that of prostatic small cell neuroendocrine carcinoma (Tai, Sun et al. 2011).
Figure 7.2 Western Blot for the expression level of proteins.

When the quantification of the expression levels was carried out in duplicates, it was interesting to find ALK expressing at very high levels in the cell line PC-3 (Figure 7.3).

The quantification of the expression levels of PPAR-γ revealed that it was expressing at high levels again in the aggressive PCa cell line- PC-3 (Figure 7.4). The highest level of
expression was however, detected in the cell line SK-OV-3, which is in line with the findings of Al-Alem et al., (2011) (Al-Alem, Southard et al. 2011).

The quantification of the expression levels of MYEOV, also very interestingly, by the average of two runs, revealed that its expression was highest in the same aggressive PCa cell line- PC-3 which is an aggressive PCa line derived from metastatic site: bone followed closely by DU-145 which is also an aggressive PCa line derived from metastatic site: brain unlike in LNCaP which is the least aggressive lymph node line among the three (Figure 7.5).
The protein expression analysis observed in the three PCa lines strengthens our hypothesis with regards the effect of ALK and MYEOV and aggressiveness of the disease. The expression of PPAR-\(\gamma\), in line with our findings of the phylogenetic tree is not dramatically over-expressed in the aggressive PCa line and further proves that obesity may be a side-effect of tobacco smoking and driving aggressiveness of the disease rather than a risk for the disease. The gene \(ALK\) comes up as one of the major confounding factors as risk of promoting aggressiveness of the disease in conjunction with \(MYEOV\). The fact that \(MYEOV\) can be regulated due to tobacco smoking is also very interesting and cannot be ignored that the protein expression of MYEOV in PC-3 was almost close to twice as in LNCaP.

7.1.4. Conclusions

Although certain risk factors for PCa are well established, yet a number of recent studies are crucial in understanding the role genetic and epigenetic make-up play for the initiation and progression of the disease (Orozco, Goh et al. 2013; Van den Broeck, Joniau et al. 2014; Vaidyanathan, Krishnamoorthy et al. 2016). Knowledge of these is useful in assisting as well
as challenging the existing risk factors. In the present study, SNP genotyping was carried out to identify SNPs and nearby genes with risk association of PCa, the demographics history of the cohort such as smoking tobacco was also analyzed for significant risk association, and finally a neighbor-joining tree-building method (Brown 2002) was used to plot and understand the effect of the genes ALK and MYEOV and tobacco smoking as a major risk for PCa better.

The findings did surprise us, more so because the gene ALK was not even considered by us in our initial SNP association study, and it being a fusion gene, perhaps promoting the alterations in cell cycle, thereby having an effect on the gene MYEOV which is an established oncogene and affecting the obesity pathway too. Further work needs to be done to define the results, but this is a good starting point analysis.
CHAPTER EIGHT

Final Discussion

"I am turned into a sort of machine for observing facts and grinding out conclusion."

- Charles Darwin
Chapter 8. Final Discussion

The research was executed with the sole aim of identifying the role of SNPs as risk for aggressive prostate cancer. We have presented herewith a multi-directional and multi-stage analysis narrowing down our total SNPs from 136 to 3.

The conclusions of our gene x environment interaction and association analyses have been the result of good planning, and thorough data mining and interpretation of results. The various challenges faced during narrowing down the list of SNPs to define them as biomarkers for aggressive prostate cancer and identifying the role of environmental factors was overcome by the final proof of concept- obtained by the protein expression analysis. It would have been interesting to see the analysed parameters and the gene x environment interactions be significantly associated in different study cohorts from New Zealand. However, that is beyond the feasibility of the current study. If our results can be reproduced, it will strengthen our findings being indicative of gene x environment interactions for PCa molecular diagnostics. We believe that chronic diseases such as cancers can be influenced by the prevalent local environmental and social factors, and should not be ignored.

Studies like these- approaching and targeting multiple frontiers to narrow down the genes of interest are very difficult to pursue not only due to the lack of well-defined data, but also the lack of proper statistical and technical approaches. The various genes identified as risk for prostate cancer (aggressive and non-aggressive) across various analyses reported in this thesis were used to plot a pathway analysis.

PCa is one of the most important male health concerns worldwide (Torre, Bray et al. 2015). There is no undermining the fact that number of deaths due to this non-skin cancer makes it one of the most crucial chronic ailments that needs to be controlled (Cooperberg, Vickers et al. 2010; Ferlay, Soerjomataram et al. 2015). New Zealand, with an unusually high
number of men with PCa, especially the aggressive kind, is on the brink of leading the tally of the percentage of men with this disease.

The three established risk factors for this disease include age, race and familial history; however, with none of these three parameters in one’s control makes the fightback even more challenging and critical. With a number of men who participated in our study having a unidirectional trend towards certain lifestyle parameters such as tobacco smoking and consuming alcohol, and high BMI, encouraged us to drive towards identifying a gene x environment interaction and the effect on certain genes on risk of aggressive PCa. This is because, the possibility of changing lifestyle and subsequently the risk of this disease seemed not a fairy tale.

We took various approaches- analyzed the gene x environment interaction while considering BMI, tobacco smoking, and alcohol consumption collectively, and adjusted the data to define the role of genes and the effect of environment on the same (Chapter 5.1). Then, we ventured into the various immunological and genetic changes occurring in an individual due to the progression of age, and defined the effect of both life-long exposure to fatty-foods and reduced exercise/ mobility, leading to high cholesterol levels and subsequently BMI changes and, tobacco smoking, alcohol consumption and effect on risk of PCa with progressing age, and the secondary effects of ageing such as lack of proper exposure to Vitamin D (Chapter 5.3, 5.4, and 5.5). The findings from these chapters did indicate that there is indeed a gene x environment interaction, and it turned us towards certain new aspects of data analysis, as we were curious to know as to what are the other possible SNPs and genes that could affect the expression and progression of PCa.

Next, we performed a machine-learning analysis and employed artificial neural networks for the same. This was a unique study on its’ own and the first time that someone had used SNP genotyping data from a case-control study, while working on aggressive PCa in New
Zealand. A set of SNPs was identified by us as risk for aggressive PCa, and, interestingly these SNPs had not been identified as statistically significant by using the standard testing tool (Chapter 6.1). When comparing the most significant SNPs between the results from the two streams of analysis, we identified only three SNPs that were significant- *MYEOV*, *LEP*, and *PPAR-γ*. This encouraged us to validate our finds by using an imported cell line.

We then identified the levels of expression of these three genes in three PCa lines (Chapter 7.1). The findings of this research was equally interesting and worrying for us. We identified very high expression levels of ALK and MYEOV- a kinase and an oncogene respectively in PC3– an aggressive PCa line. With LEP and PPAR-γ both related to fat metabolism and diabetes mellitus, which in turn has a masking effect on the diagnosis of the cancer, it will be interesting to see what happens to the expression of the genes beyond the protein expression studies. On searching the databases such as Oncomine (browsed on 22/11/2017), no uploaded datasets relating *MYEOV* to risk of aggressive PCa was found, and only one research article was identified on searching through PubMed (browsed on 23/11/2017), using the keywords “MYEOV” and “prostate cancer”. This is very promising for our research in the near future, as we believe our findings are very unique and novel and we may be the first to report these findings in a case-control study pertaining to prostate cancer.

One of the major drawbacks, however, of this research has been that we could not validate the results of gene expression with mimics of environmental factors using various cell lines, and to genotype the same, however, we did not want to influence our approach with pre-existing knowledge pertaining to these. Further work will be done with these regards and will be looked into by using various gene knockdown and knockout studies.

Since the statistical power calculation was done ad hoc by us, we will also look further into the other statistical tools to identify any other means to decide the power of an
observational case-control study as ours. This will help other researchers working in this field.

Having said the same, our strength, undoubtedly, has been the volume of our sample, the “out-of-box” approaches and executions of plans, be it the use of ANNs in analysing the SNP genotyping data or even the use of the multi-directional approach described in chapter 7.1.
"...and miles to go before I sleep, and miles to go before I sleep."

- Robert Frost
Chapter 9. Future Directions and application of work

Chapter preamble

A good research not only helps in identifying a strong direction with the core idea, but also defines the way the research will help other similar kinds of work and be developed as a major research method. Section 9.1, has been compiled as a Research Letter.

The main essence of this research was to standardize a protocol and improve upon the existing SNP biomarkers for PCa, and define some for aggressive PCa. However, if our research can be used to venture across other chronic diseases and create awareness about how important it is to consider a wide range of local parameters while diagnosing any other chronic disease. It is impossible to work in seclusion and yet carry-out path-breaking research. Our results, we believe, are pointing towards a good story, but much more needs to be done in order to prove the theories thus initiated during the course of this research. We are confident that it is a very good practical approach that was defined by us through this research.

We also intend to carry out various gene expression studies and gene knockdown and knockout studies and see the effect of the genes of interest on the progression of PCa using various publicly available online databases such as Oncomine.

Section 9.1 has been presented as a published article:

**Aggressive prostate cancer incidence in New Zealand- “united we fall, divided we stand”**

Venkatesh Vaidyanathan, Nishi Karunasinghe, Vetrivhel Krishnamurthy, Chi Hsiu-Juei Kao, Vijay Naidu, Radha Pallati, Alice Wang, Khanh Tran, Prasanna Kallingappa, Anower Jabeed, Syed M. Shahid, Jonathan Masters, Clare Wall, Ajit Narayanan, Lynnette R. Ferguson
9.1. Aggressive prostate cancer incidence in New Zealand—
‘united we fall, divided we stand’

Abstract: Prostate cancer is an important health burden to the healthcare system of any
country. However, with the current prostate-specific antigen biomarker having low predictive
value even for diagnostic purposes, the challenge is still open to tackle this chronic disease.
There have been a number of studies which have indicated and encouraged a multi-
directional approach to combat this disease. We have been carrying out a multi-directional
approach in order to identify certain New Zealand-specific factors which may be drivers for
this cancer and its aggressive forms. These will be explained in further detail in this research
letter.

Prostate cancer (PCa) is one of the most significant non-skin cancer male health concerns
worldwide (Vaidyanathan, Naidu et al. 2017), with at least 1 in 6 PCa patient estimated at
being at risk of developing aggressive PCa. This makes the identification of a strong
predictive biomarker and/or treatment of this disease a priority, especially from the New
Zealand perspective. Australia/New Zealand region records the highest rates for age-
standardised men with PCa, relative to the population of men worldwide (GLOBOCAN-
IARC 2012; Ministry of Health 2016).

Although age, ethnicity, and family history are among the most widely accepted risk
factors for PCa, nothing concrete has yet been achieved to clinically alter the outcome
(Vaidyanathan, Naidu et al. 2017). Other basic underlying components that connect these
three factors remain lifestyle and nutrition. With progressing age, lifestyle changes; different
individuals across various ethnicities enjoy different kinds of lifestyle; and certain families
also have very personalized lifestyle factors, such as the amount and kind of meat eaten.
Environmental factors play a major role in the expression of genes and the encoded proteins.
Hence, work was started in identifying the most relevant external conditions in New Zealand

There are certain environmental, nutritional and lifestyle conditions prevalent in New Zealand, such as low levels of selenium in soil (Hewitt and Dymond 2013), deficiency of Vitamin D (Ministry of Health 2012), high intake of fatty foods (University-of-Otago-and-Ministry-of-Health. 2011) and rate of obesity (Ministry-of-Health 2015), high percentage of tobacco smokers(Ministry-of-Health 2015), and ageing population (Cornwall and Davey 2004) that may combine in as yet unknown ways to increase the risk of aggressive PCa locally. We have been undertaking a holistic approach to understand the gene by environment interaction(s) and the risk of aggressive PCa in a cohort including New Zealand men of self-declared European ethnicity with different clinically diagnosed grades/stages of PCa, and gender matched healthy controls within a similar age range (Ethics reference NTY05/06/037 by Northern B Ethics Committee, New Zealand, previously, Northern Y Ethics Committee, New Zealand).

Our results have identified a number of single nucleotide polymorphisms (SNPs) statistically significantly associated with a risk of PCa and aggressive PCa. SNPs are increasingly becoming strong biomarker candidates to identify susceptibility of PCa (among other cancers). Very interestingly, a number of these genes are related directly and/or indirectly to selenium metabolism, Vitamin D metabolism, obesity and fat metabolism, inflammation and inflammatory pathways, metabolism of tobacco constituents as well as being involved with androgen metabolism, mismatch repair and oncogenesis. PCa is a common but complex disease, involving a number of aspects of genetics such as failure of mismatch repair genes and over-expression of oncogenes, but it will be naïve to forget about the impact of external factors. Current research focus is on the identification of potential and universal biomarkers
for aggressive PCa. But it is also well established that we are what we eat, and local external factors such as consumption of red meat, duration of exercise, and consumption of dietary supplements will need to be examined. This will aid us in understanding how the progression of PCa can be checked; especially bearing in mind the prevalent health and lifestyle factors in New Zealand.

Although genome-wide association studies are used for the identification of the direct role SNP association plays as risk for aggressive PCa (Turner, Kader et al. 2012), and the various environmental conditions mentioned above have also been related to various non-communicable health diseases (Moosmann and Behl 2004; Holick 2007; Bhaskaran, Douglas et al. 2014), our results indicate that SNP interactions with demographic and lifestyle factors could also add to the allelic effect of producing a modified risk of a disease (Karunasinghe, Han et al. 2012; Karunasinghe, Han et al. 2013; Vaidyanathan, Naidu et al. 2017; Vaidyanathan, Naidu et al. 2017). Those SNPs that have come up statistically significantly associated with the risk of aggressive PCa in our studies could be indicating a unique situation for New Zealand men with PCa (Vaidyanathan, Naidu et al. 2017). Our belief now is that a uniform multifactor approach will add value towards current clinical practices in improving diagnosis and along with detailed patient history is vital for combatting certain cases of aggressive PCa, which may be influenced by region-specific lifestyle factors as well as of universal genetic factors. In other words, some SNPs important for the progression of PCa may be triggered by local conditions. It is possible that local conditions also play a part for other chronic diseases as well.

The nature of PCa onset is being unraveled with further development of techniques for genomic analysis, with greater access through affordability and accuracy being key drivers of this trend. We propose that the model of patient health should unite the nature and nurture of pathologies in patients equally, and thus the risk of cancers, including PCa, should be region-
specific rather than global to take into account local external factors. By identifying such local factors, preventative education programs can also be started to help reduce the risk of PCa as well as encourage early diagnosis of PCa before it becomes aggressive. Such programs may vary from population to population taking into account lifestyle and nutritional differences. To conclude, we believe that physicians, nutritionists and dieticians, researchers, geneticists, and statisticians should be “united” in their approach to tackling PCa, which is to discuss and follow preventative measures on a local, “divided” basis.
"Learning gives creativity, creativity leads to thinking, thinking provides knowledge, knowledge makes you great."

Dr. A. P. J. Abdul Kalam
Chapter 10. Bibliography


Definition of an older or elderly person., World Health Organization.


Karunasinghe, N., Y. Zhu, et al. (2016). "Quality of life effects of androgen deprivation therapy in a prostate cancer cohort in New Zealand: can we minimize effects using a stratification based on the aldo-keto reductase family 1, member C3 rs12529 gene polymorphism?" BMC Urology 16(48).


Radhakrishnan, P., V. Chachadi, et al. (2011). "TNFalpha enhances the motility and invasiveness of prostatic cancer cells by stimulating the expression of selective


"There are no extra pieces in the universe. Everyone is here because he or she has a place to fill, and every piece must fit itself into big jigsaw puzzle."

- Deepak Chopra
### Chapter 11. Appendices

#### 11.1. SUPPLEMENTARY TABLE

Supplementary Table 1: List of SNPs associated with risk of prostate cancer selected for the current study

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Reason for inclusion in this study</th>
<th>Chromosomal Location</th>
<th>SNP ID</th>
<th>Tested Allele</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEP15</td>
<td>15 kDa selenoprotein</td>
<td><em>SEP15</em> encodes a selenoprotein, which contains a selenocysteine residue at its active site. This gene is localized on a genetic locus commonly mutated or deleted in human cancers.</td>
<td>1p22.3</td>
<td>rs5845</td>
<td>C</td>
<td>(Karunasinghe, Han et al. 2012)</td>
</tr>
<tr>
<td>PTGS2</td>
<td>Prostaglandin endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)</td>
<td><em>PTGS2</em> is responsible for the production of inflammatory prostaglandins. In cancer cells, <em>PTGS2</em> is a key step in the production of prostaglandin E2 (PGE2), which plays important roles in modulating motility, proliferation and resistance to apoptosis.</td>
<td>1q25.2-q25.3</td>
<td>rs1204276</td>
<td>G</td>
<td>(Salinas, Kwon et al. 2010)</td>
</tr>
<tr>
<td>MDM4</td>
<td>Double minute 4 protein</td>
<td><em>MDM4</em> is involved in the p53 pathway</td>
<td>1q32</td>
<td>rs1380576</td>
<td>C</td>
<td>(Sun, Lee et al. 2010)</td>
</tr>
<tr>
<td>GGCX</td>
<td>Gamma-glutamyl carboxylase</td>
<td>The 2p region is a “gene-rich region” and the SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>2p12</td>
<td>rs1018742</td>
<td>C</td>
<td>(Kote-Jarai, Olama et al.)</td>
</tr>
<tr>
<td>Gene</td>
<td>Function</td>
<td>Description</td>
<td>Chromosome</td>
<td>SNP Reference</td>
<td>Risk Allele</td>
<td>Studies</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------</td>
<td>------------</td>
<td>----------------</td>
<td>-------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td><strong>EHBP1</strong></td>
<td>EH domain binding protein 1</td>
<td>The single nucleotide polymorphism in this gene is associated with aggressive prostate cancer. Alternate splicing also results in multiple transcript variants.</td>
<td>2p15</td>
<td>rs721048</td>
<td>A</td>
<td>(Gudmundsson, Sulem et al.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs2710647 C</td>
<td></td>
<td></td>
<td></td>
<td>(Koutros, Berndt et al. 2013)</td>
</tr>
<tr>
<td><strong>THADA</strong></td>
<td>Thyroid adenoma associated</td>
<td>SNPs in this region are associated with type 2 diabetes and obesity.</td>
<td>2p21</td>
<td>rs1465618</td>
<td>A</td>
<td>(Eeles, Kote-Jarai et al. 2009)</td>
</tr>
<tr>
<td><strong>SRD5A2</strong></td>
<td>Steroid-5-alpha-reductase, alpha polypeptide 2</td>
<td><strong>SRD5A2</strong> encodes a microsomal protein expressed at high levels in androgen-sensitive tissues such as the prostate.</td>
<td>2p23.1</td>
<td>rs632148</td>
<td>C</td>
<td>(Beesley, Jordan et al. 2007; Peters, Saare et al. 2010)</td>
</tr>
<tr>
<td><strong>ITGA6</strong></td>
<td>Integrin α-6</td>
<td>The strongest SNP in the <strong>ITGA6</strong> locus, (rs12621278) has previously been reported to be associated with a 2.4-fold increased risk of prostate cancer progression.</td>
<td>2q31</td>
<td>rs1262127</td>
<td>G</td>
<td>(Cheng, Plummer et al. 2010; Lindstrom, Schumacher et al. 2011)</td>
</tr>
<tr>
<td><strong>CYP1B1</strong></td>
<td>Cytochrome P450, family 1, subfamily B, polypeptide 1</td>
<td>The gene <strong>CYP1B1</strong> encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monoxygenases which catalyze many reactions involved in the synthesis of cholesterol, steroids and other lipids.</td>
<td>2p21–p22</td>
<td>rs1056827</td>
<td>T</td>
<td>(Beuten, Gelfond et al.; Nock, Tang et al. 2007)</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Genetic Association</td>
<td>Chromosome</td>
<td>SNP</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>------------</td>
<td>-------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>MLPH</td>
<td>melanophilin</td>
<td>It has been reported that a higher melanophilin level in prostate tissue of patients with a favourable prostate cancer risk profile points out a tumour-suppressive effect.</td>
<td>2q37.2</td>
<td>rs2292884</td>
<td>(Schumacher, Berndt et al. 2011)</td>
<td></td>
</tr>
<tr>
<td>FYCO1</td>
<td>FYVE and coiled-coil domain containing 1</td>
<td>The single nucleotide polymorphism was associated with a risk for prostate cancer.</td>
<td>3p21.3</td>
<td>rs1545985</td>
<td>(Cheng, Plummer et al. 2010)</td>
<td></td>
</tr>
<tr>
<td>GPX1</td>
<td>glutathione peroxidase 1</td>
<td>Various studies have suggested anticarcinogenic effects of selenium which are probably mediated through cellular protective and redox properties of selenoenzymes <em>in vivo</em>. Certain results support a role of selenium and polymorphisms in selenoenzymes in prostate cancer etiology.</td>
<td>3p21.3</td>
<td>rs1050450</td>
<td>(Men, Zhang et al. 2014)</td>
<td></td>
</tr>
<tr>
<td>MLH1</td>
<td>mutL homolog 1</td>
<td><em>MLH1</em> is a mismatch repair gene, which encodes proteins with roles in damage recognition.</td>
<td>3p22</td>
<td>rs1799977</td>
<td>(Damaraju, Murray et al. 2006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rs9852810</td>
<td>(Langeberg, Kwon et al. 2010)</td>
<td></td>
</tr>
<tr>
<td>PPARG</td>
<td>peroxisome proliferator-activated receptor gamma</td>
<td>This gene encodes a member of the peroxisome proliferator-activated receptor (PPAR) subfamily of the</td>
<td>3p25</td>
<td>rs1779369</td>
<td>(Eeles, Olama et al. 2013)</td>
<td></td>
</tr>
</tbody>
</table>
nuclear receptors. PPAR-γ, involved in the MAPK signalling pathway, has been implicated in the pathology of a number of disease conditions such as obesity, diabetes mellitus and certain cancers.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function</th>
<th>SNP</th>
<th>Chromosome</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEFSEC</td>
<td>eukaryotic elongation factor, selenocysteine-tRNA-specific</td>
<td>rs4135275</td>
<td>3q21.3</td>
<td>(Eeles, Olama et al. 2013)</td>
</tr>
<tr>
<td>ZBTB38</td>
<td>ZBTB38 represses the transcription of methylated templates, and may therefore affect expression of certain genes</td>
<td>rs1093485</td>
<td>3q23</td>
<td>(Gudmundsson, Sulem et al.)</td>
</tr>
<tr>
<td>KIAA121</td>
<td>The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs6763931</td>
<td>4q12</td>
<td>(Eeles, Olama et al. 2013)</td>
</tr>
<tr>
<td>KIAA121</td>
<td>The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs629242</td>
<td>4q22</td>
<td>(Cheng, Plummer et al. 2010)</td>
</tr>
<tr>
<td>PDLIM5</td>
<td>PDZ and LIM domain 5</td>
<td>rs1250042</td>
<td>4q22</td>
<td>(Cooper, McGuire et al. 2011)</td>
</tr>
<tr>
<td>TET2</td>
<td>This gene is involved in the transcription of androgen receptor regulated genes KLK2 and KLK3 superpath.</td>
<td>rs7679673</td>
<td>4q24</td>
<td>(Koutros, Berndt et al.)</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Information</td>
<td>Chorosome</td>
<td>SNP</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td>FGF10</td>
<td>The protein encoded by the gene <em>FGF10</em> is a member of the fibroblast growth factor family, which possess broad mitogenic and cell survival activities, and is involved in a number of biological activities such as cell growth, tissue repair and tumour growth.</td>
<td>5p13-p12</td>
<td>rs2121875</td>
<td>G</td>
</tr>
<tr>
<td>TERT</td>
<td>The gene <em>TERT</em> is a ribonucleoprotein polymerase which maintains the telomere ends by adding telomere repeat TTAGGG. Deregulation of the expression of this gene in somatic cells may be involved in oncogenesis.</td>
<td>5p15.33</td>
<td>rs2242652</td>
<td>C</td>
</tr>
<tr>
<td>BHMT</td>
<td>The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>5q14.1</td>
<td>rs3733890</td>
<td>A</td>
</tr>
<tr>
<td>SEPP1</td>
<td>Certain results support a role of selenium and polymorphisms in selenoenzymes in prostate cancer etiology.</td>
<td>5q31</td>
<td>rs3877899</td>
<td>A</td>
</tr>
<tr>
<td>CCHCR1</td>
<td>The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>6p21.3</td>
<td>rs130067</td>
<td>A</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Chromosome</td>
<td>SNP</td>
<td>Allele</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td>-----------------</td>
<td>--------</td>
</tr>
<tr>
<td>SOD2</td>
<td>Superoxide dismutase 2 is a member of the iron/manganese superoxide dismutase family. Polymorphisms in this gene have been associated with premature aging, and certain cancers.</td>
<td>6q25</td>
<td>rs4880</td>
<td>C</td>
</tr>
<tr>
<td>JAZF1</td>
<td>JAZF zinc finger 1 The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>7p15.2-p15.1</td>
<td>rs1048656</td>
<td>7 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7p15.3</td>
<td>rs1215517</td>
<td>2 A</td>
</tr>
<tr>
<td>LMTK2</td>
<td>Lemur tyrosine kinase 2 The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>7q21</td>
<td>rs6465657</td>
<td>C</td>
</tr>
<tr>
<td>PODXL</td>
<td>Podocalyx in-like protein 1 The gene PODXL is involved in the regulation of both adhesion and cell morphology and cancer progression. May lead to increased activities of the MAPK and PI3K pathways in cancer cells.</td>
<td>7q32</td>
<td>rs3735035</td>
<td>C</td>
</tr>
<tr>
<td>LEP</td>
<td>Leptin Leptin is believed to have a role in energy homeostasis and obesity.</td>
<td>7q33</td>
<td>rs1024432</td>
<td>9 T</td>
</tr>
<tr>
<td></td>
<td>Most of the 484 annotated genes located on the 8p chromosomal region are highly likely to be oncogenes and/or tumour suppressor genes.</td>
<td>8p21</td>
<td>rs1512268</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8p21</td>
<td>rs445114</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs620861</td>
<td>A</td>
</tr>
<tr>
<td>Gene</td>
<td>rsID</td>
<td>Allele</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>rs6470494</td>
<td>A</td>
<td>(Yeager, Chatterjee et al. 2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6470517</td>
<td>A</td>
<td>(Pal, Xi et al. 2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6983267</td>
<td>A</td>
<td>(Papanikolopoulos, Landt et al. 2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1008690 8</td>
<td>C</td>
<td>(Hui, Xu et al. 2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1016343</td>
<td>C</td>
<td>(Hui, Xu et al. 2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1325473 8</td>
<td>C</td>
<td>(Fletcher, Johnson et al. 2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1378897</td>
<td>C</td>
<td>(Yeager, Chatterjee et al. 2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1690197 9</td>
<td>C</td>
<td>(Tan, Zeigler-Johnson et al. 2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs4242382</td>
<td>C</td>
<td>(Chen, Fann et al. 2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1447295</td>
<td>G</td>
<td>(Severi, Hayes et al. 2007)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6983561</td>
<td>A</td>
<td>(Suzuki, Liu et al. 2011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1690209 4</td>
<td>G</td>
<td>(Barry, Moore et al. 2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1016342</td>
<td>T</td>
<td>(Pal, Xi et al. 2009)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SLC25A37**

Solute carrier family 25 (mitochondrial iron transporter), member 37

8p21.2

rs2928679 | T | (Eeles, Kote-Jarai et al. 2009) |

8q24

- Genome-wide association studies have successfully identified a number of common polymorphisms that are strongly associated with many health conditions. A gene desert located on the chromosomal region 8q24 is associated with various types of cancers.
<table>
<thead>
<tr>
<th><strong>Gene</strong></th>
<th><strong>Protein</strong></th>
<th><strong>Chromosomal Location</strong></th>
<th><strong>SNP</strong></th>
<th><strong>Allele</strong></th>
<th><strong>Reference</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSCA</strong></td>
<td>prostate stem cell antigen</td>
<td>8q24.2</td>
<td>rs2294008</td>
<td>C</td>
<td>(Zheng, Chen et al. 2015)</td>
</tr>
<tr>
<td><strong>TLR4</strong></td>
<td>toll-like receptor 4</td>
<td>9q33.1</td>
<td>rs11536889</td>
<td>C</td>
<td>(Mansur, Gruben et al. 2014)</td>
</tr>
<tr>
<td><strong>AKR1C3</strong></td>
<td>Aldo-Keto Reductase Family 1, Member C3</td>
<td>10p15-p14</td>
<td>rs12529</td>
<td>C</td>
<td>(Yu, Huang et al. 2013)</td>
</tr>
<tr>
<td><strong>NCOA4</strong></td>
<td>nuclear receptor coactivator 4</td>
<td>10q11.2</td>
<td>rs10740051</td>
<td>A</td>
<td>(Wang, Ray et al. 2011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs17021918</td>
<td>A</td>
<td>(Hui, Wang et al. 2012)</td>
</tr>
<tr>
<td><strong>MSMB</strong></td>
<td>microsemi noprotein, beta-</td>
<td>10q11</td>
<td>rs10993994</td>
<td>C</td>
<td>(Whitaker, Kote-Jarai et al. 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs7920517</td>
<td>A</td>
<td>(Huang, Huang et al. 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs10896469</td>
<td>A</td>
<td>(Collins, Berry et al. 2005)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs17178655</td>
<td>A</td>
<td>(Sullivan, Kopp et al. 2015)</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>SNP Location</td>
<td>Allele</td>
<td>Ref.</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------</td>
<td>--------</td>
<td>-----------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>PTEN</strong></td>
<td>phosphatase and tensin homolog</td>
<td>10q23</td>
<td></td>
<td>(Xie, Lu et al. 2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>PTEN</em> is recognized as a tumour suppressor gene which is identified in a number of cancers at high frequency.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CYP17A1</strong></td>
<td>cytochrome P450, family 17, subfamily A, polypeptide 1</td>
<td>10q24.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. It has a key enzyme in the steroidogenic pathway that produces progestins, mineralocorticoids, glucocorticoids, androgens, and oestrogens.</td>
<td>rs6162</td>
<td>A</td>
<td>(Levesque, Huang et al. 2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CTBP2</strong></td>
<td>C-terminal binding protein 2</td>
<td>10q26.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs4962416</td>
<td>C</td>
<td>(Lange, Salinas et al. 2012)</td>
<td></td>
</tr>
<tr>
<td><strong>CAT</strong></td>
<td>catalase</td>
<td>11p13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>CAT</em> influences transcription factor binding, and have been identified for role in oxidative stress and diseases such as diabetes mellitus.</td>
<td>rs1001179</td>
<td>C</td>
<td>(Geybels, van den Brandt et al. 2015)</td>
<td></td>
</tr>
<tr>
<td><strong>TH</strong></td>
<td>tyrosine hydroxylase</td>
<td>11p15.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The SNP was identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs7127900</td>
<td>A</td>
<td>(Zhang, Xu et al. 2012)</td>
<td></td>
</tr>
<tr>
<td><strong>FADS2</strong></td>
<td>fatty acid desaturase 2</td>
<td>11q12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>FADS2</em> (Fatty Acid Desaturase 2) is a protein coding gene. Among its related</td>
<td>rs1535</td>
<td>A</td>
<td>(Zhang, Jia et al. 2014)</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Chromosome Location</td>
<td>SNP ID</td>
<td>Allele</td>
<td>Reference</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>--------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>FADS3</td>
<td>fatty acid desaturase 3 (Fatty Acid Desaturase 3) is a protein coding gene. This gene is clustered with family members FADS1 and FADS2.</td>
<td>11q12-q13.1</td>
<td>rs1000778</td>
<td>C</td>
<td>(Eeles, Olama et al. 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs174448</td>
<td>A</td>
<td>(Eeles, Olama et al. 2013)</td>
</tr>
<tr>
<td>MYRF</td>
<td>myelin regulatory factor</td>
<td>11q12-q13.1</td>
<td>rs174537</td>
<td>G</td>
<td>(Azrad, Zhang et al. 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1089644</td>
<td>G</td>
<td>(Zheng, Stevens et al. 2009)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs983085</td>
<td>G</td>
<td>(Sun, Purcell et al. 2008)</td>
</tr>
<tr>
<td>MYEOV</td>
<td>myeloma overexpressed</td>
<td>11q13.2</td>
<td>rs1089643</td>
<td>A</td>
<td>(Chung, Boland et al. 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs7931342</td>
<td>G</td>
<td>(Zhang, Xu et al. 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1122856</td>
<td>A</td>
<td>(Chung, Boland et al. 2012)</td>
</tr>
<tr>
<td>VDR</td>
<td>vitamin D receptor</td>
<td>12q12-q14</td>
<td>rs1116831</td>
<td>A</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs3782905</td>
<td>C</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs6823</td>
<td>C</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs7299460</td>
<td>C</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td>Gene</td>
<td>Protein Name</td>
<td>SNP Information</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUBA1C</td>
<td>tubulin, alpha 1c</td>
<td>shows sequence similarity to the steroid and thyroid hormone receptors.</td>
<td>(Holt, Kwon et al. 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRT8</td>
<td>keratin 8</td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>(Kote-Jarai, Olama et al.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDM2</td>
<td>Mouse double minute 2 homolog</td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>(Schumacher, Berndt et al. 2011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGF1</td>
<td>insulin-like growth factor 1 (somatomedin C)</td>
<td>The protein encoded by this gene is similar to insulin in function and structure and is a member of a family of proteins involved in mediating growth and development.</td>
<td>(Chang, Pao et al. 2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEPS1</td>
<td>selenoprotein S</td>
<td>Certain results support a role of selenium and polymorphisms in selenoenzymes in prostate cancer etiology.</td>
<td>(Karunasinghe, Han et al. 2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIMP</td>
<td>VCP-interacting membrane protein</td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>(Karunasinghe, Han et al. 2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USP7</td>
<td>ubiquitin specific peptidase 7</td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>(Sun, Lee et al. 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTO</td>
<td>fat mass and obesity associated</td>
<td>This gene plays an important role in fat metabolism, a major risk for aggressive prostate cancer.</td>
<td>(Lewis, Murad et al. 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELAC2</td>
<td>elaC ribonuclease Z 2</td>
<td>The SNPs were identified as a risk factor for prostate cancer.</td>
<td>(Beuten, Gelfond et al. 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>SNP</td>
<td>Chromosome</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
<td>-----</td>
<td>------------</td>
<td>------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>HNF1B</td>
<td>HNF1 homeobox B</td>
<td>rs174575</td>
<td>17p12</td>
<td>(Eeles, Olama et al. 2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA1</td>
<td>breast cancer 1, early onset</td>
<td>rs1016990</td>
<td>17q12</td>
<td>(Berndt, Sampson et al. 2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BRCA1 is an oncogene well established to have a role in breast cancer (familial).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs799923</td>
<td>A</td>
<td>(Douglas, Levin et al. 2007)</td>
<td></td>
</tr>
<tr>
<td>FASN</td>
<td>fatty acid synthase</td>
<td>rs3737559</td>
<td>17q21.31</td>
<td>(Douglas, Levin et al. 2007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>This gene plays an important role in fat metabolism, a major risk for aggressive prostate cancer.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs1859962</td>
<td>A</td>
<td>(Zhang, Cowper-Sal lari et al. 2012)</td>
<td></td>
</tr>
<tr>
<td>GPX4</td>
<td>glutathione peroxidase 4</td>
<td>rs6502051</td>
<td>17q25.3</td>
<td>(Chavarro, Kenfield et al. 2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The SNPs were identified as a risk factor for prostate cancer by a genome-wide association study.</td>
<td>rs1127678</td>
<td>A</td>
<td>(Nguyen, Ma et al. 2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs7652331</td>
<td>C</td>
<td>(Cheng, Plummer et al. 2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs713041</td>
<td>C</td>
<td>(Karunasinghe, Han et al. 2012)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs3094509</td>
<td>C</td>
<td>(Berndt, Sampson et al. 2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs3760511</td>
<td>C</td>
<td>(Nikolic, Brankovic et al. 2014)</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Chromosome</td>
<td>SNP</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>------------</td>
<td>-----</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td><strong>SLC26A6</strong></td>
<td>Solute carrier family 26 member 6</td>
<td>19q13</td>
<td>rs4794758</td>
<td>(Berndt, Sampson et al. 2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs7405696</td>
<td>(Berndt, Sampson et al. 2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs7501939</td>
<td>(Nikolic, Brankovic et al. 2014)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Solute carrier family 26 member 6</strong></td>
<td>19q13</td>
<td>rs266849</td>
<td>(Ahn, Berndt et al. 2008)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs266870</td>
<td>(Ahn, Berndt et al. 2008)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1506684</td>
<td>(Ahn, Berndt et al. 2008)</td>
<td></td>
</tr>
<tr>
<td><strong>PPP1R14A</strong></td>
<td>Protein phosphatase 1, regulatory subunit 14A</td>
<td>19q13.1</td>
<td>rs887391</td>
<td>(Hsu, Sun et al. 2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs8102476</td>
<td>(Agalliu, Wang et al. 2013)</td>
<td></td>
</tr>
<tr>
<td><strong>KLK2</strong></td>
<td>Kallikrein-related peptidase 2</td>
<td>19q13.33</td>
<td>rs1354774</td>
<td>(Jin, Zheng et al. 2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs2659122</td>
<td>(Eeles, Olama et al. 2013)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1698792</td>
<td>(Jin, Zheng et al.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs2735839</td>
<td>(He, Gu et al. 2014)</td>
<td></td>
</tr>
<tr>
<td><strong>KLK15</strong></td>
<td>Kallikrein-related peptidase 15</td>
<td>19q13.4</td>
<td>rs2659056</td>
<td>(Batra, Lose et al. 2011)</td>
<td></td>
</tr>
<tr>
<td><strong>KLK3</strong></td>
<td>Kallikrein-related peptidase 3</td>
<td>19q13</td>
<td>rs1763254</td>
<td>(Parikh, Wang et al. 2011)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs1058205</td>
<td>(Ahn, Berndt et al. 2008; Stegeman, Amankwah et al. 2015)</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Chromosome</td>
<td>SNP</td>
<td>Allele</td>
<td>Reference</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------</td>
<td>------------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>MMP9</td>
<td>Matrix metallopeptidase 9</td>
<td>20q12-13</td>
<td>rs17576</td>
<td>G</td>
<td>(Sun, Yang et al. 2011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs3787268</td>
<td>G</td>
<td>(Ting, Chen et al. 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs12793759</td>
<td>G</td>
<td>(Chung, Boland et al. 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20q13.12</td>
<td>rs3918256</td>
<td>G</td>
<td>(Jacobs, Hsing et al. 2008; Aalinkeel, Nair et al. 2011)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs2762939</td>
<td>C</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs3787557</td>
<td>C</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs4809960</td>
<td>C</td>
<td>(Holt, Kwon et al. 2010)</td>
</tr>
<tr>
<td>NUDT10/11</td>
<td>Nudix (Nucleoside Diphosphate Linked Moiety X)-Type Motif 10/11</td>
<td>Xp11.22</td>
<td>rs5945619</td>
<td>T</td>
<td>(FitzGerald, Kwon et al. 2009)</td>
</tr>
<tr>
<td>NUDT10/11</td>
<td>Nudix (Nucleoside Diphosphate Linked Moiety X)-Type Motif 10/11</td>
<td>Xp11.22</td>
<td>rs5945572</td>
<td>G</td>
<td>(Li and Gu 2015)</td>
</tr>
<tr>
<td>NUDT10/11</td>
<td>Nudix (Nucleoside Diphosphate Linked Moiety X)-Type Motif 10/11</td>
<td>Xp11.22</td>
<td>rs5919432</td>
<td>G</td>
<td>(Kote-Jarai, Olama et al.)</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td>---</td>
<td>----------------------------</td>
</tr>
</tbody>
</table>

**Table legends:**

- SNPs removed after checking for Hardy Weinberg Equilibrium using PLINK
- SNPs removed after checking for Linkage Disequilibrium using PLINK
11.2. SUPPLEMENTARY IMAGES OF Typer 4.0 MAPS
Figures 1-10: Random Sequenom TyperAnalyzer 4.0 maps for the SNPs analysed
11.3. SUPPLEMENTARY FIGURE- LD Map

Figure 11: SNPs identified to be in linkage
11.4. Co-authorship Forms

The various co-authorship forms for each published article are herewith attached.
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5, section 1 has been published as


| Nature of contribution by PhD candidate | Planned and did parts of the experiment, wrote the manuscript, did the data cleaning and statistical analysis, edited and proof-read the manuscript. |
| Extent of contribution by PhD candidate (%) | 65% |

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Naidu</td>
<td>Conceived the idea for the results section (statistics), did the data cleaning and statistical analysis, and proof-read the manuscript.</td>
</tr>
<tr>
<td>Chi Hsiu-Juei Kao</td>
<td>Did parts of the experiment and did the graphical representations, and proof-read the manuscript.</td>
</tr>
<tr>
<td>Nishi Karunasinghe</td>
<td>Planned the experiment, did parts of the experiment, and conceived the idea of the discussion chapters, and proof-read the manuscript.</td>
</tr>
<tr>
<td>Karen S. Bishop</td>
<td>Planned the experiment and did parts of the experiment, and proof-read the manuscript.</td>
</tr>
<tr>
<td>Alice Wang</td>
<td>Did parts of the experiment, and proof-read the manuscript.</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td>Conceived the idea of the discussion chapters, and proof-read the manuscript.</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

❖ the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
❖ that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Naidu</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Chi Hsiu-Juei Kao</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Nishi Karunasinghe</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Karen S. Bishop</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Alice Wang</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td></td>
<td>13/04/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5, section 1 has been published as

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Planned and did parts of the experiment, wrote the manuscript, did the data cleaning and statistical analysis, edited and proof-read the manuscript.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>65%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phillip Shepherd</td>
<td>Did parts of the experiment and proof-read the manuscript.</td>
</tr>
<tr>
<td>Jonathan Masters</td>
<td>Carried out patient recruitment and proof-read the manuscript.</td>
</tr>
<tr>
<td>Shuotun Zhu</td>
<td>Managed the database and proof-read the manuscript.</td>
</tr>
<tr>
<td>Megan Goudie</td>
<td>Carried out patient recruitment and proof-read the manuscript.</td>
</tr>
<tr>
<td>Mohanraj Krishnan</td>
<td>Conceived the idea for the results section (statistics) and proof-read the manuscript.</td>
</tr>
<tr>
<td>Anower Jabed</td>
<td>Conceived the idea of the discussion chapters and proof-read the manuscript.</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phillip Shepherd</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Jonathan Masters</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Shuotun Zhu</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Megan Goudie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mohanraj Krishnan</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Anower Jabed</td>
<td></td>
<td>20/04/2017</td>
</tr>
</tbody>
</table>
Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

**Chapter 5, section 1 has been published as**


<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Planned and did parts of the experiment, wrote the manuscript, did the data cleaning and statistical analysis, edited and proof-read the manuscript.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>65%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gareth Marlow</td>
<td>Conceived the idea of the discussion chapters and proof-read the manuscript.</td>
</tr>
<tr>
<td>Ajit Narayanan</td>
<td>Conceived the idea for the results section (statistics), did the data cleaning and statistical analysis and proof-read the manuscript.</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td>Conceived the idea of the discussion chapters and proof-read the manuscript.</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gareth Marlow</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Ajit Narayanan</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5 section 2 has been published as


Nature of contribution
by PhD candidate Conceived the idea and wrote the manuscript, edited and proof-read the manuscript.

Extent of contribution
by PhD candidate (%) 90%

CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynnette R. Ferguson</td>
<td>Conceived the idea of the discussion chapter and proof-read the manuscript.</td>
</tr>
</tbody>
</table>

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynnette R. Ferguson</td>
<td>[Signature]</td>
<td>13/04/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5 section 2 has been published as


<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Conceived the idea and wrote the manuscript, edited and proof-read the manuscript.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>90%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynnette R. Ferguson</td>
<td>Conceived the idea of the discussion chapter and proof-read the manuscript.</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynnette R. Ferguson</td>
<td>![Signature]</td>
<td>13/04/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5 section 3 is under peer review as


<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Wrote the manuscript. Did the data cleaning, data interpretation, conceived the idea for the results and discussion sections and edited and proofread the manuscript.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>80%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Naidu</td>
<td>Did the data cleaning and statistical analysis, helped conceive the idea for the results and discussion sections and proofread the manuscript</td>
</tr>
<tr>
<td>Nishi Karunasinghe</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Chi Hsiu-Juei Kao</td>
<td>Did the graphical representations and proofread the manuscript</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Anower Jabe</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Gareth Marlow</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Naidu</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Nishi Karunasinghe</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Chi Hsiu-Juei Kao</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Anower Jabe</td>
<td></td>
<td>20/04/2017</td>
</tr>
<tr>
<td>Gareth Marlow</td>
<td></td>
<td>13/04/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5 section 3 is under peer review as

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Wrote the manuscript. Did the data cleaning, data interpretation, conceived the idea for the results and discussion sections and edited and proofread the manuscript.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>80%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prasanna Kallingappa</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prasanna Kallingappa</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 5 section 4 has been conditionally accepted for publication as


<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Extent of contribution by PhD candidate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrote the manuscript. Did the data cleaning, data interpretation, conceived the idea for the results and discussion sections and edited and proofread the manuscript.</td>
<td>90%</td>
</tr>
</tbody>
</table>

CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Naidu</td>
<td>Planned and did the statistical analysis, and proofread the manuscript.</td>
</tr>
<tr>
<td>Nishi Karunasinghe</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Anower Javed</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Gareth Marlow</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td>Helped conceive the idea of the discussion section and proofread the manuscript</td>
</tr>
</tbody>
</table>

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vijay Naidu</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Nishi Karunasinghe</td>
<td></td>
<td>19/04/2017</td>
</tr>
<tr>
<td>Anower Javed</td>
<td></td>
<td>20/04/2017</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Gareth Marlow</td>
<td></td>
<td>13/04/2017</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td></td>
<td>13/04/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 9 section 1 has been accepted for publication as Vaidyanathan V, Karunasinghe N, Krishnamurthy V, Kao CH, Naidu V, et al. Aggressive prostate cancer incidence in New Zealand- “united we fall, divided we stand”

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Extent of contribution by PhD candidate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceived the idea and wrote the manuscript.</td>
<td>65%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nishi Karunasinghe</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Vetrivhel Krishnamurthy</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Chi Hsiu-Juei Kao</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Vijay Naidu</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Alice Wang</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nishi Karunasinghe</td>
<td>Signature</td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Vetrivhel Krishnamurthy</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Chi Hsiu-Juei Kao</td>
<td>Signature</td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Vijay Naidu</td>
<td>Signature</td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Radha Pallati</td>
<td>Signature</td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Alice Wang</td>
<td>Signature</td>
<td>27/11/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 9 section 1 has been accepted for publication as
Vaidyanathan V, Karunasinghe N, Krishnamurthy V, Kao CH, Naidu V, et al. Aggressive prostate cancer incidence in New Zealand- “united we fall, divided we stand”

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Extent of contribution by PhD candidate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceived the idea and wrote the manuscript.</td>
<td>65%</td>
</tr>
</tbody>
</table>

CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khanh Tran</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Prasanna Kalligappa</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Anower Jabed</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Syed M. Shahid</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Jonathan Masters</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Clare Wall</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
</tbody>
</table>

Certification by Co-Authors

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khanh Tran</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Prasanna Kallingappa</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Anower Jabed</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Syed M. Shahid</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Jonathan Masters</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Clare Wall</td>
<td></td>
<td>27/11/2017</td>
</tr>
</tbody>
</table>
This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 9 section 1 has been accepted for publication as

Vaidyanathan V, Karunasinghe N, Krishnamurthy V, Kao CH, Naidu V, et al. Aggressive prostate cancer incidence in New Zealand- "united we fall, divided we stand"

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Conceived the idea and wrote the manuscript.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>65%</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajit Narayanan</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td>Proofread the manuscript and suggested corrections</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajit Narayanan</td>
<td></td>
<td>27/11/2017</td>
</tr>
<tr>
<td>Lynnette R. Ferguson</td>
<td></td>
<td>27/11/2017</td>
</tr>
</tbody>
</table>